NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

# Predictive Analytics applied to Alzheimer's Disease: A Data Visualisation Framework for Understanding Current Research and Future Challenges

Hugo Alexandre Martins Esteves

Dissertation as a partial requirement for obtaining the master's degree in information management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

**NOVA Information Management School**
**Instituto Superior de Estatística e Gestão de Informação**
Universidade Nova de Lisboa

# PREDICTIVE ANALYTICS APPLIED TO ALZHEIMER'S DISEASE: A DATA VISUALIZATION FRAMEWORK FOR UNDERSTANDING CURRENT RESEARCH AND FUTURE CHALLENGES

by

Hugo Alexandre Martins Esteves

Dissertation as a partial requirement for obtaining a master's degree in information management, with a specialisation in Business Intelligence and Knowledge Management.

**Advisor:** *Prof. Mauro Castelli*

November 2018

# DEDICATION

"What would life be if we had no courage to attempt anything.", Vincent Van Gogh

*To my grandfather.*

# ACKNOWLEDGEMENTS

Firstly, my sincere and honest gratefulness to Professor Mauro Castelli, from Nova IMS for believing in this work's idea; for continually believing in me; for never letting go this project until its completion; for his availability for taking this project forward; for all the support provided during the ups and downs; for all his honest, insightful and valuable contributions and knowledge sharing.

My honest thank you for Professor Pedro Cabral, from Nova IMS, for all the guidance during the idea conceptualisation, who helped me shaping and reshaping the overall idea of the work despite a moment of uncertainty, and who boosted my confidence in an early stage for carrying on with this contribution for the scientific community.

My strong appreciation for Eng. João Leal and Eng. Bruno Valente, from José de Mello Saúde, for their valuable insights about the reality in the private sector, for their availability for active collaboration, and all the genuine help provided on trying to get reliable medical data

To Ana Esteves for her knowledge sharing about the state of the art of data visualisation techniques and publications, which quickly became the inspiration foundation for part of the developed work.

Lastly but not least, my profound thank you to Andreia, whose unconditional love made all this happen.

# ABSTRACT

Big Data is, nowadays, regarded as a tool for improving the healthcare sector in many areas, such as in its economic side, by trying to search for operational efficiency gaps, and in personalised treatment, by selecting the best drug for the patient, for instance. Data science can play a key role in identifying diseases in an early stage, or even when there are no signs of it, track its progress, quickly identify the efficacy of treatments and suggest alternative ones. Therefore, the prevention side of healthcare can be enhanced with the usage of state-of-the-art predictive big data analytics and machine learning methods, integrating the available, complex, heterogeneous, yet sparse, data from multiple sources, towards a better disease and pathology patterns identification. It can be applied for the diagnostic challenging neurodegenerative disorders; the identification of the patterns that trigger those disorders can make possible to identify more risk factors, biomarkers, in every human being. With that, we can improve the effectiveness of the medical interventions, helping people to stay healthy and active for a longer period. In this work, a review of the state of science about predictive big data analytics is done, concerning its application to Alzheimer's Disease early diagnosis. It is done by searching and summarising the scientific articles published in respectable online sources, putting together all the information that is spread out in the world wide web, with the goal of enhancing knowledge management and collaboration practices about the topic. Furthermore, an interactive data visualisation tool to better manage and identify the scientific articles is develop, delivering, in this way, a holistic visual overview of the developments done in the important field of Alzheimer's Disease diagnosis.



# KEYWORDS

Big Data; Alzheimer's disease; Predictive Modelling; Predictive analytics; Data Mining; Decision Support Systems; Data Visualization; Knowledge Management; Information Systems; Information Management

# RESUMO

*Big Data* é hoje considerada uma ferramenta para melhorar o sector da saúde em muitas áreas, tais como na sua vertente mais económica, tentando encontrar lacunas de eficiência operacional, e no tratamento personalizado, selecionando o melhor medicamento para o paciente, por exemplo. A ciência de dados pode desempenhar um papel fundamental na identificação de doenças em um estágio inicial, ou mesmo quando não há sinais dela, acompanhar o seu progresso, identificar rapidamente a eficácia dos tratamentos indicados ao paciente e sugerir alternativas. Portanto, o lado preventivo dos cuidados de saúde pode ser bastante melhorado com o uso de métodos avançados de análise preditiva com *big data* e de *machine learning*, integrando os dados disponíveis, geralmente complexos, heterogéneos e esparsos provenientes de múltiplas fontes, para uma melhor identificação de padrões patológicos e da doença. Estes métodos podem ser aplicados nas doenças neurodegenerativas que ainda são um grande desafio no seu diagnóstico; a identificação dos padrões que desencadeiam esses distúrbios pode possibilitar a identificação de mais fatores de risco, biomarcadores, em todo e qualquer ser humano. Com isso, podemos melhorar a eficácia das intervenções médicas, ajudando as pessoas a permanecerem saudáveis e ativas por um período mais longo. Neste trabalho, é feita uma revisão do estado da arte sobre a análise preditiva com *big data*, no que diz respeito à sua aplicação ao diagnóstico precoce da Doença de Alzheimer. Isto foi realizado através da pesquisa exaustiva e resumo de um grande número de artigos científicos publicados em fontes online de referência na área, reunindo a informação que está amplamente espalhada na *world wide web*, com o objetivo de aprimorar a gestão do conhecimento e as práticas de colaboração sobre o tema. Além disso, uma ferramenta interativa de visualização de dados para melhor gerir e identificar os artigos científicos foi desenvolvida, fornecendo, desta forma, uma visão holística dos avanços científico feitos no importante campo do diagnóstico da Doença de Alzheimer.

# PALAVRAS-CHAVE

Big Data; Doença de Alzheimer; Modelação e análise preditiva; *data mining*; sistemas de suporte à decisão; visualização de dados; gestão do conhecimento; sistemas de informação; gestão de informação

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ACRONYMS AND ABBREVIATIONS

| | |
|---|---|
| AD | Alzheimer's Disease |
| AD-PS | Alzheimer's Disease Pattern Similarity |
| ADAS-Cog | Ad Assessment Cognitive Scale |
| ADNI | Disease Neuroimaging Initiative Database |
| AVLT | Auditory Verbal Learning Test |
| BOPEL | Bayesian Outcome Prediction with Ensemble Learning |
| CaMCCo | Cascaded Multiview Canonical Correlation |
| CDR | Clinical Dementia Rating |
| CVLT-LDTR | California Verbal Learning Test Long Delayed Total Recall |
| DARTEL | Diffeomorphic Anatomical Registration Through Exponentiated Lie Algebra |
| DSI | Disease State Index |
| DSI | Statistical Disease State Index |
| DT | Decision Tree |
| EMR | Electronic Medical Records |
| EN-RLR | Elastic Net Regularized Logistic Regression |
| ESR | Erythrocyte Sedimentation Rate |
| ESTI | Stands for The Parameter Estimation Error |
| GDS | Geriatric Depression Scale |
| GP-LR | Bayesian Gaussian Process Logistic Regression |
| GR-NN | General Regression Neural Network |
| GRNN | Generalized Regression Neural Network |
| HGM-FS | High-Order Graph Matching Based Feature Selection |
| iMSF | Multi-Source Feature Learning Method |
| iSFS | Incomplete Source Feature Selection |
| LDA | Linear Discriminant Analysis |
| LDS | Low-Density Separation |
| LLE | Locally Linear Embedding |
| LR | Logistic Regression |
| MCI | Mild Cognitive Impairment |
| MH | Modified Hachinski |
| MKL | Multiple Kernel Learning |
| MMSE | Mini-Mental State Examination |
| mRMR | Minimum Redundancy Maximum Relevance |
| MSE | Mean Squared Error |
| NB | Neuropsychological Battery |
| NIQ | Neuropsychiatric Inventory Questionnaire |
| NPI-Q | Neuropsychiatric Inventory Questionnaire |
| NPSE | Neuro-Psychological Status Exams |
| OPLS | Orthogonal Partial Least Squares to Latent Structures |
| PET | Positron Emitting Tomography |
| PMCI | Progressive Mild Cognitive Impairment |
| PPV | Positive Predictive Value |
| RAVLT | Rey Auditory Verbal Learning Test |

| | |
|---|---|
| rDAm | Randomized Denoising Autoencoder Marker |
| RF | Random Forest |
| RLR | Regularized Logistic Regression |
| RMSE | Root Mean Square Prediction Error |
| SMCI | Stable Mild Cognitive Impairment |
| SMML | Sparse Multimodal Learning |
| SMOTE | Synthetic Minority Over-Sampling Technique |
| SNP | Genetic Modality Single-Nucleotide Polymorphism |
| SOM | Self-Organising Maps |
| SVD | Singular Value Decomposition |
| WHIMS-MRI | Women's Health Initiative Magnetic Resonance Imaging Study |
| WRAP | Wisconsin Registry For Alzheimer’s Prevention |
| ZARADEMP | Zaragoza Dementia And Depression Project |

# 1. INTRODUCTION

There are more than 47 million people that have from dementia worldwide at the present moment (Prince, Comas-Herrera, Knapp, Guerchet, & Karagiannidou, 2016). The forecasts tell us that the number will double by 2030 and more than triple by 2050, reaching more than 131 million by that year, as populations ages. Moreover, the diagnostic coverage is low among people that live with dementia, being between 40% to 50% worldwide, even in most high-income countries, whereas less than 10% of cases are diagnosed in low and middle-income countries, (Prince et al., 2016).

The most common cause of dementia is Alzheimer's Disease (AD), which accounts for an estimated 60 to 80 per cent of cases (Prince et al., 2016). About half of these cases involve Alzheimer's pathology solely; many have evidence of pathologic changes related to other dementias. Moreover, between the year 2000 and 2014, deaths resulting from stroke, heart disease, and prostate cancer decreased 21%, 14%, and 9%, respectively, whereas deaths from AD increased 89% during that period (Association, 2017).

Another aspect that is worth mentioning is the enormous neurodegenerative diseases' economic impact, being the US $818 billion the estimated worldwide cost of dementia and expected to become a trillion dollar disease by 2018 (Prince et al., 2016), exceeding the combined current market values of Apple and Google. We already know that dementia is bringing tremendous costs for the society, at many levels (Wang, Kung, & Byrd, 2015; Wang, Kung, Wang, & Cegielski, 2017), as the burden is increasing with the population ageing in the most developed countries. In fact, in 2016, more than 15 million family members and other unpaid caregivers provided an estimated 18.2 billion hours of care to people with Alzheimer's or other dementias, which accounts to an estimated unpaid work of more than $230 billion (Association, 2017).

In recent years there have been intense efforts to develop and validate AD biomarkers, including those detectable with brain imaging, in the blood and using cerebrospinal fluid, have intensified. Such efforts will transform the practice of diagnosing AD from cognitive and functional symptoms focused to a biomarker-based diagnosis. This new approach promotes an earlier stage accurate diagnosis of disease and potentially leads to a more accurate understanding of AD prevalence and incidence. At the same time there has been intense focus on neurodegenerative pathologies and on Alzheimer's Disease, not only by the research community but also by the most critical stakeholders in the world's healthcare ecosystem, such as universities, research institutions, pharma industry, venture capital investment firms and philanthropic foundations, such as Bill Gates Foundation (Bill Gates, 2017).

Although the Electronic Health Records (EHR) have been long seen as one import missing element for unlocking the potential of deeper research areas in medical practice (Prather et al., 1997), in which correlations and patterns identification could provide more in-depth knowledge, the accumulation of massive quantities of data and information about patients and their medical conditions in clinical databases has only started recently. With that, the need to work that data rose and, consequently, data science started to be more known and even became a discipline in computer sciences. From there to the "big data" concept was only a matter of a couple of years, being accelerated, in particular, in the biomedical area, due to the decoding and complete sequencing of the human genome. In fact, "big data", which is considered a hype for some due to its daily banalization in the IT-driven industries, has gained much attention recently, at the same pace as the EHRs, particularly in areas of biomedicine where society still face clear unmet medical needs. One can easily find "big data" associated with a new tool for complex

problem solving (Hofmann-Apitius, 2015), big data approaches seem to open promising perspectives for a better understanding of complex neurodegenerative diseases. Therefore, big data principles and techniques have been thought to have tremendous potential in the dementia-related research. We live in a world where technologies and the needed computational power are already available for that; those might unlock many possibilities for knowledge discovery. With a quick search in the www, one can easily find that healthcare practices are becoming more and more empowered through data science, machine learning and artificial intelligence.

Following that "big data promise", many studies have been addressing Alzheimer's disease diagnosis, prognosis, predicting its evolution in patients, and many other relevant scientific breakthroughs using data science methods and using the available research datasets. This wealth of information that comes out of the scientific researches is becoming increasingly difficult to follow, jeopardising the potential of a more collaborative approach for incremental innovation.

Therefore, with so many advances, articles and scientific papers published, a newcomer to the field might ask why we still haven't found all the disease patterns that would lead to a cure or the delay in the disease evolution in each patient. What if we could predict that someone will develop AD way before the symptoms even show up, with some degree of certainty? Perhaps we are still far behind that goal, but the latest scientific developments are helping us to discover hidden patterns in the data, either from genomics, proteomics, and many more, data of all kinds, to get there. Firstly, we need to tackle the problems of an accurate and timely diagnosis, which has been a real struggle. It is a path that we must walk with a collaborative and incremental innovative mindset. Thus, it is of the utmost necessity that we build integrative data and information management systems, that can act as the foundation for the research to take the most advantage of the quantity of health data that we collect.

This dissertation is a small step towards that end. Information was gathered and organised from many scientific publications about the application of predictive analytics to the Alzheimer's Disease discovery, trying to tap the question of what has been done so far, what methods and datasets were used, which are the ones that were not tried yet, and so on. The answer could lead to the conclusion that it is not straightforward to gather information about what was already done by the scientific community concerning neurodegenerative diseases. However, we might ask instead whether the scientific community is working towards an integrative approach to the datasets available and to the methodologies used to extract knowledge from it. Considering the latest scientific publications in the field can we say that we are, as a society, working to truly unlock the big data analytics potential for a more accurate, and applicable, prediction of AD's development in each human being, even before the symptoms' manifestation? Have we already tested all the predictive models, datasets, biomarkers, feature selection methods, to name a few, that we could think of?

These questions are addressed in this work through its three main objectives : firstly through an in-depth scientific literature review for the identification of the relevant science in the area; then by the extraction of the relevant data science-related content from the scientific publications found, such as the predictive models used in them; and then the development of a framework for an interactive data visualization solution to enhance the comprehension of those answers.

Furthermore, this work aims to address the gap of how difficult it is to gather the correct information for a compelling visualization of what was already done and what is still to be done; it also tackles a need in the scientific community of having integrative and truly collaborative information management systems,

by gathering and summarising all the scientific articles stored in specific online individualised databases about the usage of predictive analytics applied to AD discovery. It is a tentative to enhance the information management practices towards better and faster collaboration systems on the topic of AD's prediction and prevention.

To achieve that, the primary goal of this dissertation is to provide an objective and complete overview of state of the art regarding the application of predictive analytics and big data analytics in the early detection of the AD. That one can find in chapter 2.

The method used for the search of relevant publications in the field and the methods used for the analysis and categorisation of data science-related content are presented in Chapter 3. In that chapter, one can find the thought process that led to the filtering and clustering of the articles found by the type of data, datasets, the feature selection methods applied, the learning and classification algorithms used or developed.

With so many contributors in the scientific community in the field of predictive analytics and machine learning applied to neurodegenerative diseases, there is the need to start converging to a common ground of knowledge gathering and sharing, to foster quicker collaboration and innovation in the field. Moreover, with so many scientific articles repositories spread out in the world wide web, the urge of making those scientific developments and publications visible in one place and in a user-friendly way becomes more significant than before. In the words of the current United Nations' Deputy Secretary-General, "we need to develop data literacy, innovative tools and data visualisation platforms, which allow users to understand data intuitively and interact seamlessly with data in real time.*".*

To help to achieve that goal, focused on a specific research topic, Alzheimer's Disease, a data visualisation framework was developed. "We visualise to communicate a point" (Evergreen, 2016)**,** to enhance our communication efficacy when complexity is involved. Data visualisation became a relevant field of computer science, not much because it is a brand-new area which requires a blend of new professional skills, but mostly because we need to communicate better the insights collected from the worked data. With data, we gather information. With information we can show the story, transforming complexity into simplicity. That will help us to create knowledge.

However, there are no tools specifically for extracting the information about what was done in all the publications and articles of a specific field of research. There are, though, many research done about scholarly data visualisation (Liu et al., 2018), but that is mostly focused on authors, researchers, institutions, publication keywords, and so on. Furthermore, that data can then be a useful source of information to advance collaboration and a better understanding of the current state of the art in that research field. This information is vital not only for the scientific community in general but also for the institutional decision-makers and the general public, tapping into the "open science" concept, in the idea of augmenting the power of the collective intelligence.

With this in mind, in Chapter 4 one can find the methodology and tools used to develop the interactive data visualisation framework, aiming to provide an up to date, intuitive and user-friendly way not only to identify relevant articles in the field but also their contribution to the application of predictive analytics methods to the Alzheimer's Disease discovery. The visualisation framework is based on the business intelligence software Microsoft Power BI Desktop, version 2.64, which provides the desired interactivity and flexibility layers to the solution.

In Chapter 5 the discussion of what has been achieved through this work is done, followed by a detailed explanation of the limitations of this work and by the conclusion. One can also find in this chapter the suggestions for future work.

Lastly, it is worth to point out that, despite following a structured approach supported by extensive literature, it is impossible to claim that all the published articles under the topic were found and reviewed. Moreover, the search process was guided by the combination of selected keywords used in predetermined online private publications databases, which provided reasonable confidence that it was possible to collect a relatively large sample of articles. It is, indeed, impossible to find all the scientific publications on a specific topic given the way the online publication market is structured.

# 2. LITERATURE REVIEW

## 2.1. **Big Data and Predictive Analytics in Healthcare**

We cannot deny it – the rise of information technologies has changed healthcare practices and industry forever. The amount of data stored in the healthcare facilities all over the world is gigantic. Clinics, hospitals, public and private research centres, and other private organisations have already accumulated large amounts of information about their patients and their medical conditions. Relationships and patterns within this data started to provide new medical knowledge. Since the beginning of the new century, methodologies have been developed and applied to discover this hidden knowledge, which started to be applied to a vast range of healthcare fields, such as obstetrics, where one could evaluate for factors potentially contributing to preterm birth (Prather et al., 1997). Since then, many developments on the area of data mining, undoubtedly the foundation of data science broad concept, were done, and all the methods necessary for accurate data modelling, descriptive and predictive analytics were as well improved and refined. Consequently, the rise of these methods allowed its application to advance healthcare practices in general and, in particular, access treatment effectiveness, healthcare management, customer relationship management, among others (Kaur & Wasan, 2006; Koh & Tan, 2005; Tomar & Agarwal, 2013).

### 2.1.1. **The Rise of Big Data**

In the last decade the rise of certain technologies, such as the Internet of Things (IoT) which lead to the proliferation of wearable sensors and the birth of the term *eHealth*, the mobile phone usage massification and the mobile electronic diagnostic systems for chronic disease symptoms tracking, which was responsible for the birth of the *mHealth* (Bhavnani, Narula, & Sengupta, 2016), have had a tremendous impact on the patient-generated data across the developed countries. With the resources available, we can now extract information and, hopefully, knowledge out of that increasing amount of big data generated. As a consequence of the rise of such data-driven approaches, clinical databases have gathered large quantities of information about patients and their medical conditions, from which relationships and patterns within this data could provide new medical knowledge. Consequently, there are an increasing number of Information and Communication Technology (ICT) companies trying to develop and evolve its systems and algorithms to become prime providers of analytical solutions for the healthcare sector, being one of the most famous ones the IBM Watson Health solution (Belle et al., 2015; Y. Chen, Argentinis, & Weber, 2016; Hoyt, Snider, Thompson, & Mantravadi, 2016; Pham, Tran, Phung, & Venkatesh, 2017; Sun & Reddy, 2013). The biggest challenge that those systems want to overcome lies in harnessing volumes of data, integrating the data from hundreds of sources, and understanding their various formats, so that advanced reasoning, predictive modelling, and machine learning techniques can be used to advance research faster.

Big data became a cornerstone of the future of healthcare. Big data in biomedicine and biotechnology is characterised by (Hofmann-Apitius, 2015) three main concepts, which are usually called *The Three Vs of big data* (Table 1): volume, velocity and variety. More recently, a fourth aspect has been added, veracity, which entails the issue of heterogenety of data quality and the need for curation. This is particularly important in medical related areas, where the quality of data (most of them are still annotations) varies tremendously.

*Table 1 – The three Vs of* big *data*

| | |
|---|---|
| **Volume** | due to the massive "data flood" we experience in biomedicine from the developed data generating technologies and scientific advances such as genome sequencing, neuroimaging and informatisation of clinical procedures records (EHR); |
| **Velocity** | not only associated with the processing of real-time data analysis but also with the heterogeneity of data as far as timescales are concerned. |
| **Variety** | a crucial step in big data analytics practices is the integration of heterogeneous data, where shared semantics for metadata annotation play an important role. |

Therefore, it is easy to understand that extracting information and knowledge from Big Data will improve the likelihood of tacking issues in healthcare more proficiently and faster, as well as evolving healthcare practices to evidence-based medicine, helping on genomic analytics, healthcare device and remote monitoring, patient profile analytics, among others (Raghupathi & Raghupathi, 2014). It has become so important that many studies have been published on the topic of big data and predictive analytics: the digitalization of healthcare (Bhavnani et al., 2016), diagnostics in healthcare from imaging (Ithapu et al., 2015), blood and cerebrospinal fluid (Association, 2017), and genomics-based analytics (Belle et al., 2015) and to many more problems in health sector (Huang et al., 2015). This is a clear evidence that a wide range of stakeholders in our society, have the desire to harness as fast as possible the data-driven technologies to address the healthcare challenges we will face in the future, being genomics, imaging and signal based analytics (Belle et al., 2015; Raghupathi & Raghupathi, 2014) the most promising big data analytics areas. These big data-based analytical areas have been applied to many difficult diseases such as breast and pancreatic cancer, diabetes, leukaemia, and many more (Belle et al., 2015).

Some years ago, the concept of "Computer-Aid Diagnosis" for management of medical big data for diagnosis purposes arose (Siuly & Zhang, 2016). According to some authors, the neurologists expect these systems to provide support in their decision-making process, helping them to how medical big data can be managed by computation intelligent system in neurological diseases diagnosis. Current diagnosis technologies, such as magnetic resonance imaging (MRI) and electroencephalogram (EEG), produce huge quantity data (in size and dimension) for detection, monitoring and treatment of neurological diseases. The next step is to have computerised diagnosis systems, called "computer-aided diagnosis (CAD)" that can automatically detect the neurological abnormalities combining all medical big data available for the patient in a holistic and integrative way. This system has the objective of improving consistency on disease diagnosis and on increasing, consequently, the success of treatment as well as on saving lives and reducing costs and time, consequently.

Furthermore, many frameworks have been proposed to uniformize the heterogeneity of the complex big healthcare data world. A Digital Health Framework (DHF) was developed with the goal to embrace the emerging requirements, in terms of data and tools, of applying systems medicine into healthcare (Cano, Tenyi, Vela, Miralles, & Roca, 2017), where there are different levels of medical data and information for

biomedical research (Figure 1). These kinds of frameworks, either specific to a particular area in medicine (Bochicchio, Cuzzocrea, & Vaira, 2016) or, instead, general ones (Sakr & Elgammal, 2016), have already been proposed and tested in real hospitals environments (Ojha & Mathur, 2016).

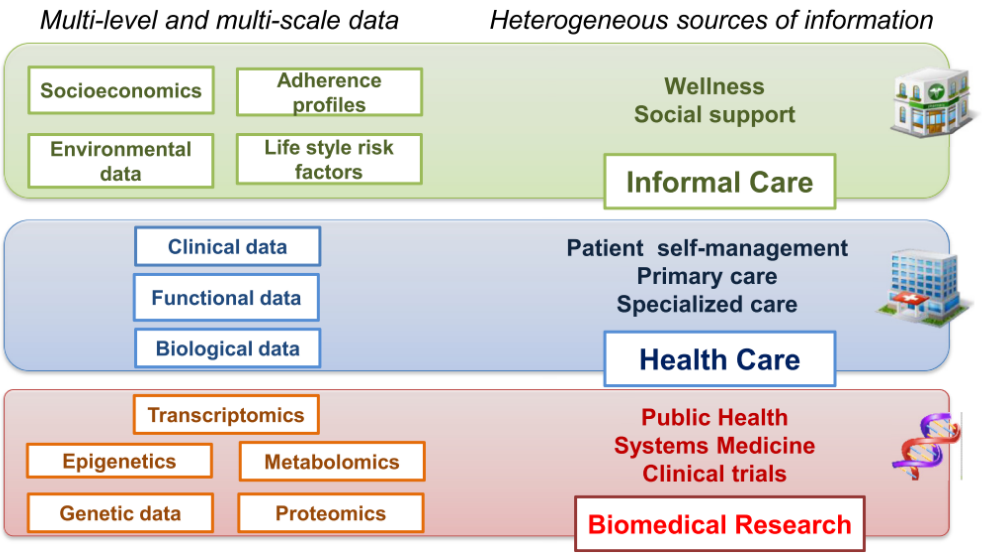


*Figure 1 –Framework with a multilevel integration of heterogeneous patient information generated by different data sources (Cano et al., 2017).*

However, like in everything in life, these Information technologies bring difficult challenges with them, which must be considered right at the development and implementation stages. Such an approach is crucial to creating trust in such systems, even more, when they are to be delivered in such sensitive area as Healthcare. The important aspect of data quality required for these systems to work properly is one of them, touching the "V" of veracity explained earlier. Other ones would be the entire lifecycle of health data, the problems arising from errors and inaccuracies in the data itself, the sources and the pedigree of data, and how the underlying purpose of data collection impact the analytic processing and knowledge expected to be derived (Sukumar, Natarajan, & Ferrell, 2015). Automation in the form of data handling, storage, entry and processing technologies should be thought as a tricky and hard balance to get. At one level, automation is what we all desire for speed and convenience, while it can create a different set of data quality issues as far as data management is concerned that can be hard to solve afterwards.

It is worth noticing that although there are many studies about the potential benefits of big data and predictive analytics implementation for healthcare organizations management in real life, financially and operationally wise (Wang & Hajli, 2017; Wang et al., 2017), some of them even developing complex deep learning methods (Pham et al., 2017), this work does not focus on those aspects. Those applications are of great importance for the future of healthcare, so much that there have been some authors that have tried to summarize all the articles written about that subject (Malik, Abdallah, & Ala'raj, 2018) and check if those approaches have been implemented in the field (Wang, Kung, & Byrd, 2015).

Once talking about big data analytics, one should try to understand which type of analytics is referred to descriptive or predictive analytics. In this work, the predictive potential of big data is the main focus, although the descriptive analytics is very common in imaging-based analytics, i.e., to detect patterns in the data generated from MRI, for instance. The scientific studies using descriptive, analytical methods –

the most commonly used methods of data mining techniques, are mainly focused on the diagnosis (Tomar & Agarwal, 2013). Predictive healthcare approach is of utmost importance and should be done in a close partnership between public and private organizations, so that not only the main goal is reached, i.e. to act in the prevention side and, consequently, prevent diseases, but also to generate and use the synergies that come from that collaboration (Wang, Kung, Ting, et al., 2015) - economic savings to the overall healthcare system can be obtained in this way.

Predictive healthcare have been applied to several fields that have been considered problematic for some of the world's future societies, one of them being the diabetics (Pham et al., 2017; Saravana Kumar, Eswari, Sampath, & Lavanya, 2015), where the diabetes types prevalent are predicted, as well as other consequences associated with it, as well as the type of treatment to be provided. These systems have the ultimate goal of providing efficient ways to cure and to care the patients, with better outcomes such as affordability and availability.

### 2.1.2. **From Big to Smart Data**

Big data is often discussed in the context of both improving medical care and also in preventing diseases as it has the power of facilitating action on the modifiable risk factors that contribute to a large fraction of the chronic disease burden, such as physical activity, diet, tobacco use, and exposure to pollution (Barrett, Humblet, Hiatt, & Adler, 2013). It can do so by facilitating the discovery of risk factors at population, subpopulation, and individual levels, and by improving the effectiveness of the intervention.

As we have already seen, the widespread use of healthcare information systems, in the form of the known EHRs, leads the way healthcare could and must be delivered, by resorting to the power of big data, making the ambition of delivering individualised healthcare a possibility in the near future. We already have the tools to say that the future of medicine will be through the personalisation of medical care, after the human genome being entirely sequenced. Big data and predictive analytics state-of-the-art algorithms, such as neural networks (Pham, Tran, Phung, & Venkatesh, 2016; Pham et al., 2017), genetic algorithms (Tejeswinee, Shomona, & Athilakshmi, 2017) and deep learning (Ravi et al., 2016), are key components of that vision, as data should be collected throughout time in order to build accurate models. Healthcare observations, recorded in Electronic Medical Records (EMR), are episodic and irregular in time. Thus, personalised, predictive medicine requires the modelling of patient illnesses and all the care processes, among other things, as depicted in Figure 1, which inherently have long-term temporal dependencies of patient data (Pham et al., 2016). The dream of these systems would be not only to model the diseases progressing and recommend accurate diagnosis but also to predict future risks for each human being, way before the symptoms appear. This is of utmost need for heavy social and economic burden diseases like diabetes and neurodegenerative diseases. All this has made, undoubtedly, data mining and predictive analytics important tools for healthcare decision making.

However, despite all the published literature and all the interest, among the scientific community, for knowledge discovery from the big data to improve the delivery of healthcare services, there has been a lack of attempts for synthesizing the available information on how the big data phenomenon has contributed already to better outcomes (Malik et al., 2018). Furthermore, only a small proportion of the huge potential of risk prediction modelling is being applied (Cano et al., 2017), in particular for health forecasting chronic patients, due to the lack of implemented procedures for accessing and mining health information from daily clinical practice. Big data and predictive analytics will only bring advantages if, and only if, the access, storage and process of quality multilevel, multisource data is possible. Only in this way,

a holistic strategy for subject-specific risk prediction can be achieved, in such a way that early identification of patient susceptibility to multi-morbidity might enable cost-effective preventive healthcare strategies and enhance management of chronic patients (Cano et al., 2017).

Nevertheless, developments on the area of information management and systems, machine learning and data science, by using advanced algorithms and analytical methods, such Gravitational Search Algorithm (Nagpal, Arora, Dey, & Shreya, 2017), are being used to extract as much knowledge as possible from the data we have gathered. However, without existing yet a quick, easy and integrative way to find them - the process of locating all scientific breakthroughs is still manual, even experienced authors in the field struggle to find all the published articles they need (Malik et al., 2018), research will necessarily be slower than they could be.

### 2.2. **Neurodegenerative Disease's Diagnosis**

One of the most difficult challenges for modern medicine is the diagnosis of neurological diseases (Siuly & Zhang, 2016). According to the World Health Organization's recent report, neurological disorders, such as epilepsy, Alzheimer's disease and stroke to a headache, affect up to 1 billion people worldwide. An estimated 6.8 million people die every year because of neurological disorders. A lot is being done by international organisations to raise the awareness of the burden society will carry in the future due to dementia and other neurodegenerative diseases. In 2012, the World Health Organization declared dementia as a public health priority, when 35.6 million people worldwide were living with dementia, according to estimates. In 2016, the number rose to 47 million, more than the population of Spain (Prince et al., 2016).

The diagnosis has always been a headache for medical practitioners, with huge social impact. Many authors have tackled the problematic of neurodegenerative diseases diagnosis (Kaur & Wasan, 2006; Siuly & Zhang, 2016), the degenerative process, its progress and treatment efficacy of some specific diseases. There has been as well several drug discovery programs (Geerts et al., 2016), mainly initiated by the pharma industry, among many other research programs and initiatives (Alzheimer's Association, 2016).

Due to the increasing awareness and focus on the neuro diseases - the cases of dementia are increasing year over year (Perera et al., 2018), many previous studies, both in animals and in humans, have examined the relationship of Parkinson's disease (PD) risk to trauma, genetics, environment, co-morbidities, or lifestyle, to find patterns and biomarkers. For AD, for instance, there has been ground-breaking research studies regarding the identifications of such potential biomarkers, such as sleep (Bubu et al., 2017), diabetes (Habes et al., 2016; Janson et al., 2004), antidepressant and anti-hypertense drugs (Habes et al., 2016), risk factors related with lifestyle such as alcohol and smoking (Habes et al., 2016; World Health Organization, 2012), as well as using structural imaging patterns of advanced brain aging, capturing aging atrophy patterns (Habes et al., 2016).

Mild Cognitive Impairment (MCI) is currently considered to be an early stage of a neurodegenerative disease. Moreover, patients diagnosed with MCI are assumed to have a higher risk to evolve to the AD. Prediction of progression from a stage of MCI to dementia is a major pursuit in current research (Lin et al., 2018; Pereira et al., 2017), whether by predicting the likelihood of MCI conversion or the time to actual conversion. These kinds of studies have often shown a relatively high statistical confidence level on the prediction results from the predictive models developed.

Although focused on one neurodegenerative disease, such as Parkinson's disease, some studies have the particularity of being possible to extend and applied to other diseases such as Alzheimer's Disease (Dinov et al., 2016), with clear research cost-savings and great impact on society. This studies usually tackle on the feature selection phase of knowledge discovery, which is a crucial step in the data mining process when dealing with such degree of data heterogeneity (Tejeswinee et al., 2017).

In practical terms, the clinical assessment in clinics and hospitals of patients with suspected dementia is done nowadays using mostly structural imaging based on magnetic resonance (MRI). Prospective data, from preclinical to other stages of Alzheimer disease, are radically changing how the disease is conceptualised and will influence its future diagnosis and treatment.

Given the complexity of neurodegenerative diseases (Hofmann-Apitius, 2015), it was expected a significant heterogeneity of data in that scientific area. Heterogeneity of data concerns their mode (omics data, imaging data and clinical data, leading to variety) as well as their quality/veracity. Big data analytics uses a wide range of data integration, modelling, and mining strategies in order to understand and predict systems behaviour in complex systems. The integration of complex and heterogeneous big data from multiple sources offers unparalleled opportunities to study the early stages of prevalent neurodegenerative processes, track their progression and quickly identify the efficacies of alternative treatments. Even in such complex data integration systems, data mining techniques have proven already to be a forceful way for biomarker discovery, with which significant findings of different clinical and biological biomarkers have been possible. However, due to the brain-related disorders complexity, challenges are still to be overcome, i.e. the way to integrate such data diversity in a single model/system in a straightforward and standardised way is still to be done. Nevertheless, the solution relies on data from different sources and model integration (data from single sources are unable to explain complex biological processes), where several data types are combined to provide complementary views (Carreiro, Mendonça, de Carvalho, & Madeira, 2015), to find and/or validate composite biomarkers using more powerful predictive models.

Another important aspect to consider is the EHRs. Huge volumes of health data are being collected and electronically stored, either in routine EHR databases or through research-driven cohort studies associated with biobanks and other efforts (Perera et al., 2018). To help improving access to patient-level data, the European Medical Information Framework (EMIF) was launched in 2013; it aims to create an environment that allows efficient reuse of health data in two therapeutic areas, one being Alzheimer's disease. One important initiative of EMIF consortium was to assemble what is at the present moment one of the world's largest EHR resource for dementia research. However, according to some authors of recently published articles, EHRs have been underused in dementia research (Perera et al., 2018; R. Zhang, Simon, & Yu, 2017), which entail a great potential for research and are a key element for the accuracy of the risk prediction of neurodegenerative diseases in healthy patients.

At the same time that biomedical and high-dimensionality datasets are becoming increasingly larger and freely accessible for neurodegenerative diseases, it is also important that new information analytics platforms be developed, allowing the big data resources to become an actionable mechanism, for faster disease prediction, a more accurate diagnosis and advanced therapeutic development. These systems will allow cross-platform datasets correlation (data of distinct types, sources or modalities). Machine learning methods are now strategies applied by the scientific community to drive the future development of diagnostic and therapeutic applications for Alzheimer's disease, such as latent semantic analytics,

topological data investigation, and deep learning techniques. However, the simplicity and diversity of platforms are needed, as the simple repetition of similar analytical methods will always be unproductive (Maudsley, Devanarayan, Martin, & Geerts, 2018)**.**

### 2.2.1. **Predictive Analytics Applied to Alzheimer's Disease Discovery**

Alzheimer's disease is considered "a slowly progressive brain disease that begins well before clinical symptoms emerge" (Alzheimer's Association, 2015). So far, the AD cannot be prevented, slowed or cured (R. Zhang et al., 2017). In 2011, an estimated 24 million people worldwide had dementia, the majority of whom were thought to have Alzheimer's disease (Ballard et al., 2011). An estimated 5.3 million Americans of all ages had Alzheimer's disease, in 2015 (Alzheimer's Association, 2015).

There are significant studies and research with the goal of having an evidence-based early diagnosis of the disease (Alzheimer's Association, 2016; R. Chen et al., 2012; Frisoni, Fox, Jack, Scheltens, & Thompson, 2010; Janson et al., 2004; McEvoy et al., 2009; Serrano-Pozo, Frosch, Masliah, & Hyman, 2011; Tomar & Agarwal, 2013) but big data can further increase the reliability of the desired early stage diagnosis and even, being the ultimate goal, anticipate the disease by extracting enough meaningful information about every individual. With the knowledge gathered from millions of patients, and with methods such as machine learning, artificial intelligence, pattern recognition, the knowledge extracted could be used to predict, with a high level of certainty, whether a person will develop AD in the future.

There is already strong evidence of potential risks and protective factors for Alzheimer's disease, dementia, and cognitive decline, but further work is needed to understand these better and to establish whether early interventions can substantially lower these risks. Then, if possible, the potential patient will have time to change its behaviour, avoid some risk factors, delay the disease appearance, or even not develop it at all in the future. Data science will help not only on the journey towards the discovery of a cure but also on helping to reduce the rather increasing growth pace of the disease's prevalence worldwide, which will put pressure on several sectors of our society, in particular, in the underdeveloped countries' economies (Association, 2017).

Neurologists, psychiatrist and geriatricians, use a variety of approaches and tools to help make a diagnosis (Association, 2017), as there is no single test for Alzheimer's. Thus, the data collected to help on that difficult diagnosis can be one or many of the following:

- Medical and family history from the individual, including psychiatric history, and history of cognitive and behavioural changes;
- Information from a family member on observed changes in thinking skills and behaviour;
- Cognitive tests, physical and neurologic examinations;
- Blood samples and brain imaging to rule out other potential causes of dementia symptoms, such as a tumour or certain vitamin deficiencies.

Although physicians can almost always determine if a person has dementia, it may be difficult to identify the exact cause. Sometimes weeks may be needed for the individual to complete the required tests and examinations, and for the physician to interpret the results to make a proper and accurate diagnosis.

### 2.2.2. **Biomarkers and Data Sources**

Biomarkers such as magnetic resonance imaging (MRI)s – perhaps the most used so far (Beheshti, Demirel, & Matsuda, 2017; Casanova et al., 2018; Moradi, Pepe, Gaser, Huttunen, & Tohka, 2015), positron emission tomography (PET), cerebrospinal fluid (CSF), and genetic modality single-nucleotide polymorphism (SNP) have been widely used in recent years for Alzheimer's disease diagnosis, revealing to be an extremely useful sources of insightful data. However, some studies go beyond this and use EEG signal processing in order to help in AD early identification (Rodrigues, Teixeira, Garrett, Alves, & Freitas, 2016).

A relatively large amount of studies have focused on how to utilise these biomarkers to classify AD, and a relatively big number of them suggest that combining them for prediction is better than using any of them independently (Z. Zhang, Huang, & Shen, 2014). Still, some authors use a single source of data already existing, such as NACC neuropsychological/neuropsychiatric datasets (Beheshti et al., 2017; Lemos et al., 2012; Lin et al., 2018) as the source for the developed machine learning technique to build, for instance, accurate MCI conversion prediction calculator (Lin et al., 2018), whereas others use their own generated dataset, e.g. neuropsychological data, to discriminate Alzheimer disease from MCI (Lemos et al., 2012). Even using single data sources, they presented great results by using sophisticated data mining and machine learning methodologies, such as ensembles of deep learning architectures (Ortiz, Munilla, Górriz, & Ramírez, 2016).

As already mentioned, structural MRIs biomarkers are quite well known and used in this field. Many studies have been carried out for the development of robust automated disease assessment algorithms (Frisoni et al., 2010). When sMRI are combined with the most common focus areas in this field - the prediction of MCI to AD conversion, usually with the goal of improving the opportunities for MCI and AD patients to get involved in clinical trials at the same time that the trial is cost-effective and efficient (Ithapu et al., 2015; Lin et al., 2018; McEvoy et al., 2009; Rodrigues et al., 2016), it results in the most common reseach topic among the published scientific articles found for this work. Within this research topic, many initiatives were started such as the development of machine learning methods (Casanova et al., 2018) and frameworks (Moradi et al., 2015), usage of Bayesian networks (R. Chen et al., 2012; Land & Schaffer, 2016), the application of deep learning (Ithapu et al., 2015) and genetic algorithms for feature selection (Beheshti et al., 2017).

With the introduction of MCI as a diagnostic category, challenges of diagnosing Alzheimer's Disease (AD) come along with it, as no single marker has been proven to categorise accurately patients into their respective diagnostic groups accurately. Thus, some recent studies tackle the gaps identified in previous ones, attempting to develop combined predictors of AD and MCI that simultaneously consider all diagnostic categories and provide optimal combined representations using different set of modalities (e.g. neurophysiologic, MRI, PET, CSF Proteomics, genotype) for the prediction (Singanamalli et al., 2017). As current datasets in medicine can have many missing values in some modalities, convolutional neural networks were also studied in a multi-modality imaging data setup for the estimation of missing modality (Li et al., 2014).

Actually, according to R. Zhang et al. (2017), all the articles about the application of big data analytics on AD discovery fit in one of these seven categories: diagnosing AD or MCI, predicting MCI to AD conversion, categorizing risks for developing AD, text mining the literature for knowledge discovery, predicting AD

progression, describing clinical care for persons with AD, and understanding the relationship between cognition and AD.

Actionable knowledge from all these studies using big data is what some researchers argue (Geerts et al., 2016). They support the idea of generating actionable knowledge that could help in developing new treatment paradigms, either for a better trial participant selection or by advancing the way we deliver medical practices to the society, including new drug discovery's efficacy.

### 2.2.3. **Information Availability and Management**

The production and availability of data in some repositories and data lakes in some countries, far from being centralised and interoperable, for dementia research will increase with the same rate as we can observe in other complex medical research areas. Therefore, the need for increased interoperability of data will simultaneously increase (Hofmann-Apitius, 2015). It will push for substantial efforts to cope not only with the rapid growth of data volume but also with the notorious lack of interoperability as far as data, information, and, first and foremost, knowledge is concerned. From the literature and the practical cases out there, data mining and machine learning techniques, such as deep learning, will more and more be applied to this field in the future, and there is no doubt that integrative modelling and mining approaches (Carreiro et al., 2015) will happen and will have a strong impact on dementia research. It is the only way forward if we want to harness the full potential data can give us for discovering hidden patterns that we could find in the past, without such technologies at our disposal.

As a consequence of the need of data and information integration, and interoperability, in the field of information management, and applied particularly to troubling growing health problems such as of Alzheimer's disease, generated interest and space for the work presented here to happen.

However, without having neither a single platform where all this information is stored nor an interoperable and interconnected system of information sharing, it becomes hard to keep track of all the developments in the field, threatening the true potential of knowledge discovery and sharing, and decreasing the power of true collaboration and incremental innovation. However, there has been recently in the business world the need for data and information integration – one can easily recall a couple of startups and companies that were created based on data centralisation, which are now considered data aggregators in industries such as travel, finance, media, and many more. Knowledge management and collaboration have common and mutually interdependent purposes and practices (Qureshi, Briggs, & Hlupic, 2006). Collaboration technologies have had huge implications in the value creation for organisations and society in a whole; it is even more critical in areas of innovation, from where many people around the world would benefit. Collaboration technologies, in the form of information technologies, unlock the benefits of collaboration, through which the intellectual capital created can be used to create value. It can, potentially, even accelerate novel drug development and, simultaneously, aid rational drug repurposing for such complex conditions such as AD (Maudsley et al., 2018). Thus, there is an urge for more subtle and intuitive informatics systems. Information is spread, and for actionable knowledge we need it to be quickly accessible.

In (R. Zhang et al., 2017), one can find a first approach for summarising all the knowledge around the topic in one single work. In (Carreiro et al., 2015), an extensive list and its detailed description of the types of data sources available and used for predictive modelling in AD research is done. One must dig in the world wide web and spend many hours to find these useful scientific publications. Furthermore, as it was

already concluded by some authors (R. Zhang et al., 2017), big data are increasingly used to address AD-related research questions, but EHR datasets are still underutilised; they should be used to further empower the potentialities of the predictive analytics for AD discovery. It is one conclusion that could be quickly drawn from a single, quick analysis of integrative and intuitive information management and visualisation platform, which, if existed, would avoid many hours, and funding, spend on trying to capture the overall picture/state of the science of that specific research topic.

Therefore, it is of vital importance that we seek, and need that we already see rising in some European forums (Tennant, 2018), of having an open and intuitive information management platform towards open science, where these studies would be visible for the entire research community and the entire world. If so, it will provide a holistic view of the relevant research topics and fostering innovation and collaboration between, for instance, distant physical cohorts, and lead to an increase in the research efficiency and effectiveness, which would, consequently, reduce the probability of the research topics' overlap. The wasted time on articles search is inefficient, taken that we live in an era where information is always at our reach. Incremental innovation methodologies based on collective intelligence should be a right mindset for faster discoveries of hidden knowledge in areas so complex as in Alzheimer's Disease.

### 2.3. **Data Visualization**

Data and information in databases are hard to see. Humans have a limitation in how much information can they understand looking to a table with data all over the place. Data visualisation is nothing more than being creative on how we visually present the data in order to help humans to deprehend concepts, patterns, insights that could not be seen and understood by looking at the same data in a spreadsheet, for instance. It is a happy marriage between visual arts, computer science and, being more precise, graphical design and data science.

### 2.4. **Visualise to Communicate**

Data visualisation became a critical field of computer science, so much that there already a world-renowned competition and award for good data visualisation (Kantar, 2018). With the rise of the data storage we have seen in the past years, that new forms of visualising the same data appeared, and with it a new industry and more and more data science professionals and becoming knowledgeable on that. All this contribute for data visualisation to become the single most effective communication enhancer for data science and related areas, given the number of books and industry publications about the topic (Munzner & Maguire, 2014). In fact, "we visualise to communicate a point" (Evergreen, 2016).

According to the author Munzner & Maguire (2014), visualisation design can be split in three main fronts: what data users need to see, why users need to carry out their tasks, and how the visual representations proposed can be constructed and manipulated. Visualisation, comprehending a big set of areas, touches as well in the realm of psychophysics, as human perception plays an essential role in the area of visualisation (Healey, 2007; Munzner & Maguire, 2014). Visualisation is much more than data in a bar chart; there are "bad data", "bad taste" and "bad perception", which can turn out to be a nightmare for the ones trying to convey the right information visually in the right way (Healy, 2018). In the same line, Edward R. Tufte is unarguably the best-known critic, by far, of information visualisation, mentions in this well-known book from 1983, *The Visual Display of Quantitative Information* (Tufte, 2001), summaries his view on how difficult it is to get graphics and data visualisation right:

*"Graphical excellence is the well-designed presentation of interesting data—a matter of substance, of statistics, and design (…) [It] consists of complex ideas communicated with clarity, precision, and efficiency. (…) [It] is that which gives the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space (…) [It] is nearly always multivariate … graphical excellence requires telling the truth about the data." (Tufte, 1983, p. 51).*

### 2.4.1. **Data Visualization for Scholarly Data**

Scholarly data visualisation in one branch of data visualisation's broad range of applications. The typical examples of scholarly data visualisation are the visualisation of researchers, papers, journals, and institutions. These are quite common, as far as on scholarly social or information network analysis is concerned. All these entities are usually represented regarding generated scholarly networks, wherein nodes represent the academic entities/researchers and links represent the relationships such as citation, co-author relationship and institution partnerships. In Liu et al. (2018) one can find extensive research about the topic, more specifically regarding the software commonly used for that purpose and the research that has been carried out.

There are visualisation techniques specially designed for those simple attributes (e.g. author, institution) and heterogeneous networks of scholarly data (Liu et al., 2018). Moreover, one initiative that tackles the need for identifying trends in the research areas is led by the *Automatic Identification of Research Trends* initiative (AIDA). It has already diverse case studies on research positioning and trend identification, which are mostly used by PhD candidates, researchers, group leaders, and policymakers. This initiative and studies, among many others, use the tool VOSviewer (Vosviewer, 2018). This tool provides the possibility to create collaboration maps, citation density maps, among other options, making use of the interconnection of articles through their citations. Other existing tools that can be used for the same objective are SCI2 (sci2, 2018), Histcite (Histcite, 2018) and Citespace (citespace101, 2016).

Although there are many studies, mostly from computer science related institutions, regarding the development of systems for the extraction of relevant metadata to build visualisation solutions, the reality is that the data of most scientific documents on the web are unstructured or semi-structured. That becomes a hurdle and justifies the proliferation of studies in this field. Automatic article's metadata extraction methods become, therefore, an important task. In Z & H (2011), for instance, it is developed a system to extract the title, authors and abstract from scientific papers based on spatial and visual knowledge principle. They argue that the system developed achieves high accuracy in header metadata extraction so automatic index creation for digital libraries can effectively be achieved with it.

Another issue to overcome in this research space if also regarding the data quality and the lack of structure in the online digital libraries, which are highly dependent on the classifier's quality used to mine that data from a set of crawled documents, which in turn depends, among other things, on the choice of the feature representation. To tackle this problem, Caragea et al. (2014) propose novel features that result in effective and efficient classification models for automatic identification of research articles, using CiteSeerx digital library for their successful experiment.

Keywords have been as well an exciting target for the ones studying ways to organise better the scholar databases and willing to come up with innovative data visualisation for a better understanding of that data. To achieve that, some researchers (Isenberg, Isenberg, Sedlmair, Chen, & Möller, 2017) have

recently collected the research paper's author-assigned keywords manually from the PDFs of all papers of the visualization research field, published in a significant number of years, for their study on the importance of establishing common visualization terminology for accurate categorization of papers. That is another proof that organising and structuring this data is not a straightforward task. Their work could, among other things, facilitate the crucial step of finding the right reviewers for research submissions. One interesting outcome of their work was a web-based search tool that allows visualisation researchers to quickly browse thousands of keywords used for their research area to find related work or make informed keyword choices.

If the metadata, in particular the keywords and the main topic of the research article, are correctly identified and stored in an online public database, cross-domain research topic mining can be done to help users find relationships among related research domains, obtaining a quick overview of these domains (Jiang & Zhang, 2016); it is another investigation topic that uses the power of simple, yet effective, data visualisation and visual analytics to help the scientific community to be more efficient and knowledgeable. That is even further enhanced if the time dimension is included in that cross-domain topic correlation. Visualisation of research topics trends in a cross-domain based approach makes total sense, in such a fast-paced world. Keeping track of what the knowledgeable authors of those topics and domains can also be useful for foster collaboration and mentorship. Tapping into that topic, Fung (2015) introduced an innovative and exciting way to visualise a bibliographic database. He presented in his study on egocentric visualisation of a bibliographic database, a visualisation design approach based on a botanic tree metaphor, resulting in a rich depiction of one's research publication records.

However, little has been done with the mapping of the relevant content of those articles. As far as the research for this work has shown and understood from the research done on the topic, no methodology or framework was developed yet for extracting the information about what data science methods were applied in all the publications and articles in a specific field of research. That would be quite feasible to do for the majority of the scientific documentation out there, as there is a reasonably limited amount of possible data mining and predictive analytical methods, and related fields. It cannot be generalised for every field of science, though. Nevertheless, with the increasing demand in our societies for faster, agile, scalable and replicable solutions for efficient resource management, this approach of mining the literature for more than just the common metadata will have to be applied.

Furthermore, that data is a source of good information about that specific topic in order to advance collaboration and enhance the understanding of the current state of the art, tapping into the global knowledge management topic. This information is vital not only for the scientific community in general but also for the institutional decision-makers, and, ultimately, for the general public. This can be the baseline for a form of information visualisation and should stand for an *open science* concept of scientific publication. It should be based on an augmented collective intelligence framework, for a concise and user-friendly understanding of the content.

# 3. METHODOLOGY

## 3.1. **Research Articles and Publication Collection**

In this chapter one can find the detailed description of the methodology that was followed to access, collect, review and organise the scientific literature about the use of predictive analytics to the Alzheimer's Disease discovery, as well as the description of the database and data visualisation platform development. The methodology followed is depicted in the flowchart of Figure 2.

### 3.1.1. **Sources Selection**

Although the topic of the search is quite limited to this work title, a quick search online (e.g. using Google Search) was enough to understand that it covers a broad range of subtopics. Therefore, the query returned thousands of articles, and a more limiting search was needed.

The scientific literature collection was carried out by using a selective and systematic approach throughout the chosen online sources. The online sources used for a systematic scientific literature search were:

- Science Direct
- Scopus
- PubMed – Medline
- Web of Science

Then, the platform called *Mendeley* was used not only to store and manage the relevant articles (in a first screening) but also to search for related documents within it (thought its connection to Elsevier's database) thus, only articles stored in Elsevier and Science Direct would be found using this method).

The article access limitations found in some online resources, in which a payment is required to read the article, have limited the number of articles gathered.

### 3.1.2. **Search Keywords Selection**

Based on the scientific literature review presented in chapter 2 and to bridge the knowledge gap, the objective was to gather as many articles as possible about the application of data science methods, in particular, the ones regarding predictive analytics and big data, to the Alzheimer's disease research. Thus, the goal was to identify publications that focused on the power of data to accelerate the research of AD, touching the perspective of predicting the development of the disease in a healthy human being.

Having those web portals as a reference, the keywords used for the scientific literature retrieval in those online sources for this work, in the form of different combinations, were:

{(Alzheimer, Alzheimer's disease, AD)

AND

(big data, data analytics, analytics, algorithm, data mining, prediction, predictive, predictive analytics)}

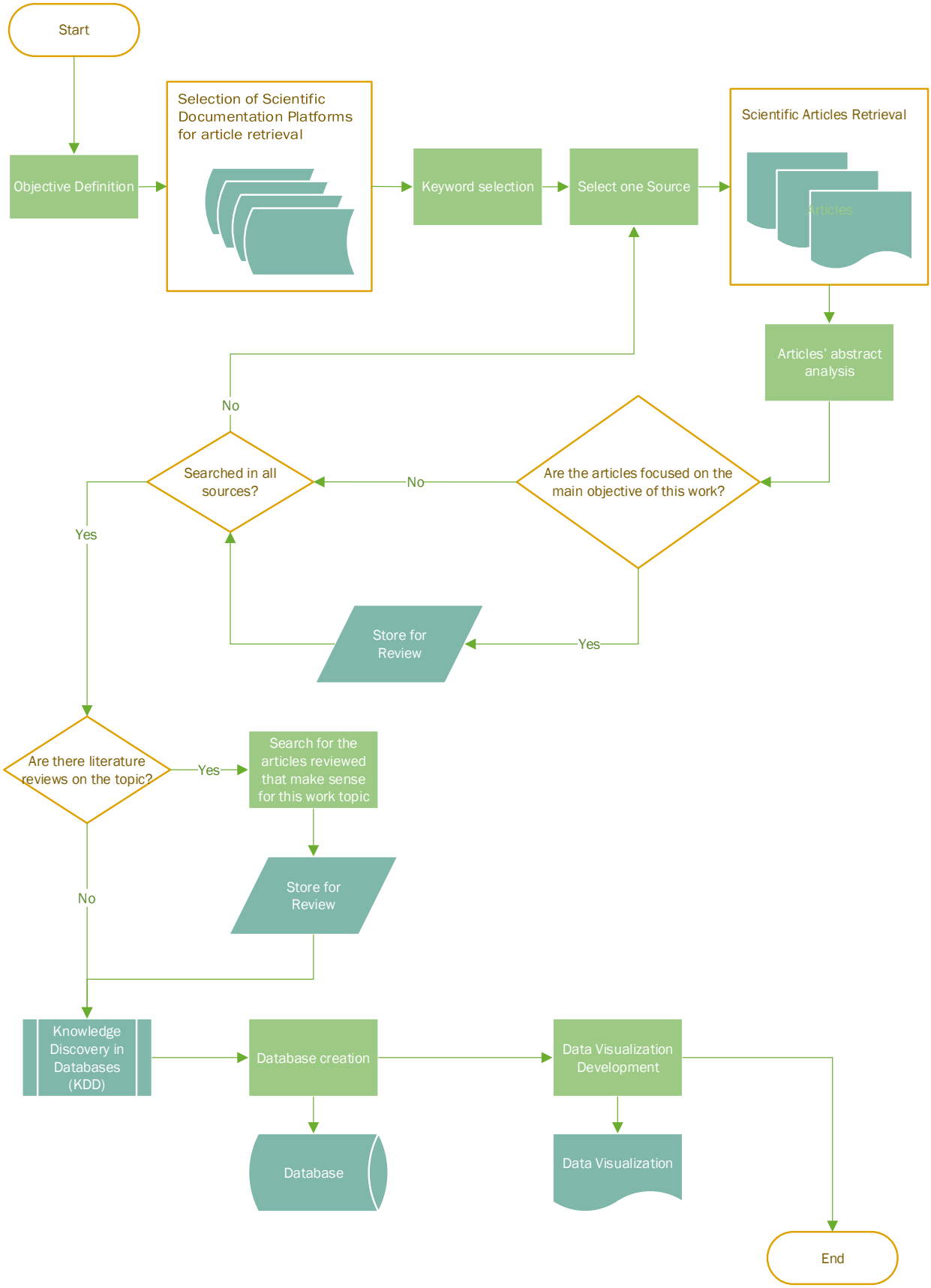


*Figure 2 – Flow chart explaining the methodology followed in this work.*

### 3.1.3. **Article Abstract Analysis**

Many research articles were also found, on purpose, about the predictive analytics application in healthcare in general, in order to build not only chapter 2 but also to understand the research evolution in data science application to health-related topics. These were not included in the database and, therefore, will not be visible in the data visualisation platform.

Regarding the time filter chosen and considered acceptable for the literature search, it was not set at all because this is a relatively new area. The narrowing down of the documentation was then applied concerning relevancy for the topic of this work. No other filters were used in the search step.

### 3.1.4. **Descriptive Analysis**

The total number of articles found were 20, and those are the ones useful for the primary objective of this thesis. However, four were theoretical, in which neither predictive data modelling nor real patient data were applied (Carreiro et al., 2015; Geerts et al., 2016; Haas et al., 2016; Maudsley et al., 2018). Therefore, only 16 were selected.

Additionally, one systematic literature review article was found (R. Zhang et al., 2017), in which a review was done on the application of big data on the AD discovery and research. Thus, and taking into consideration that specific literature review article had some objectives in common with this work, the articles reviewed in the former were also considered to be reviewed *a posteriori*; if the main topic of the study in each of those reviewed articles fit the primary objective of this work, these would be considered.

The result was that 36 articles were considered and included in this work, after a complete content analysis. In summary, 58 articles were included in the database, and the complete list can be found in the Annex section, being 20 the ones found in the online resources.

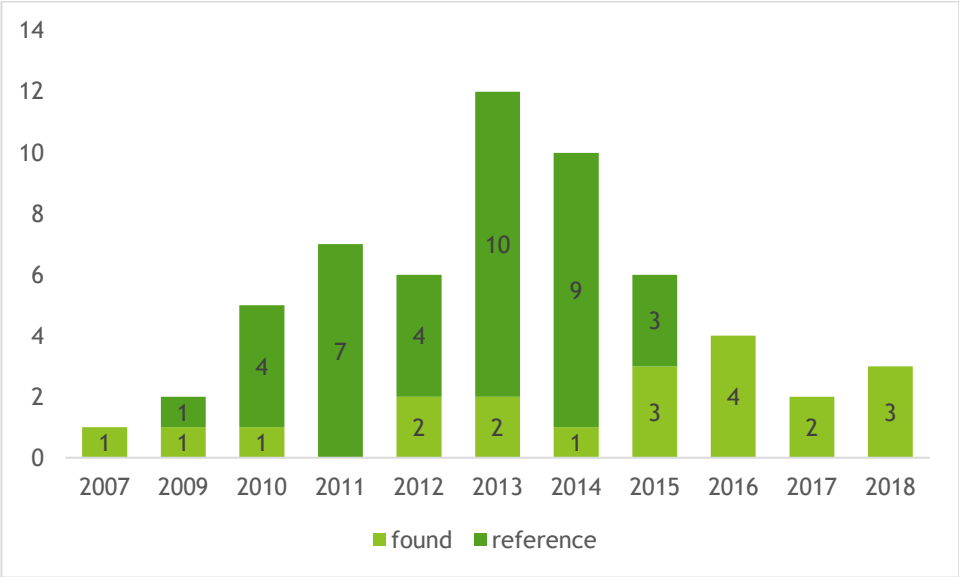


*Figure 3 – Total number of scientific articles found and included in the database, either by using the search criteria or through reference articles, per year of publication.*

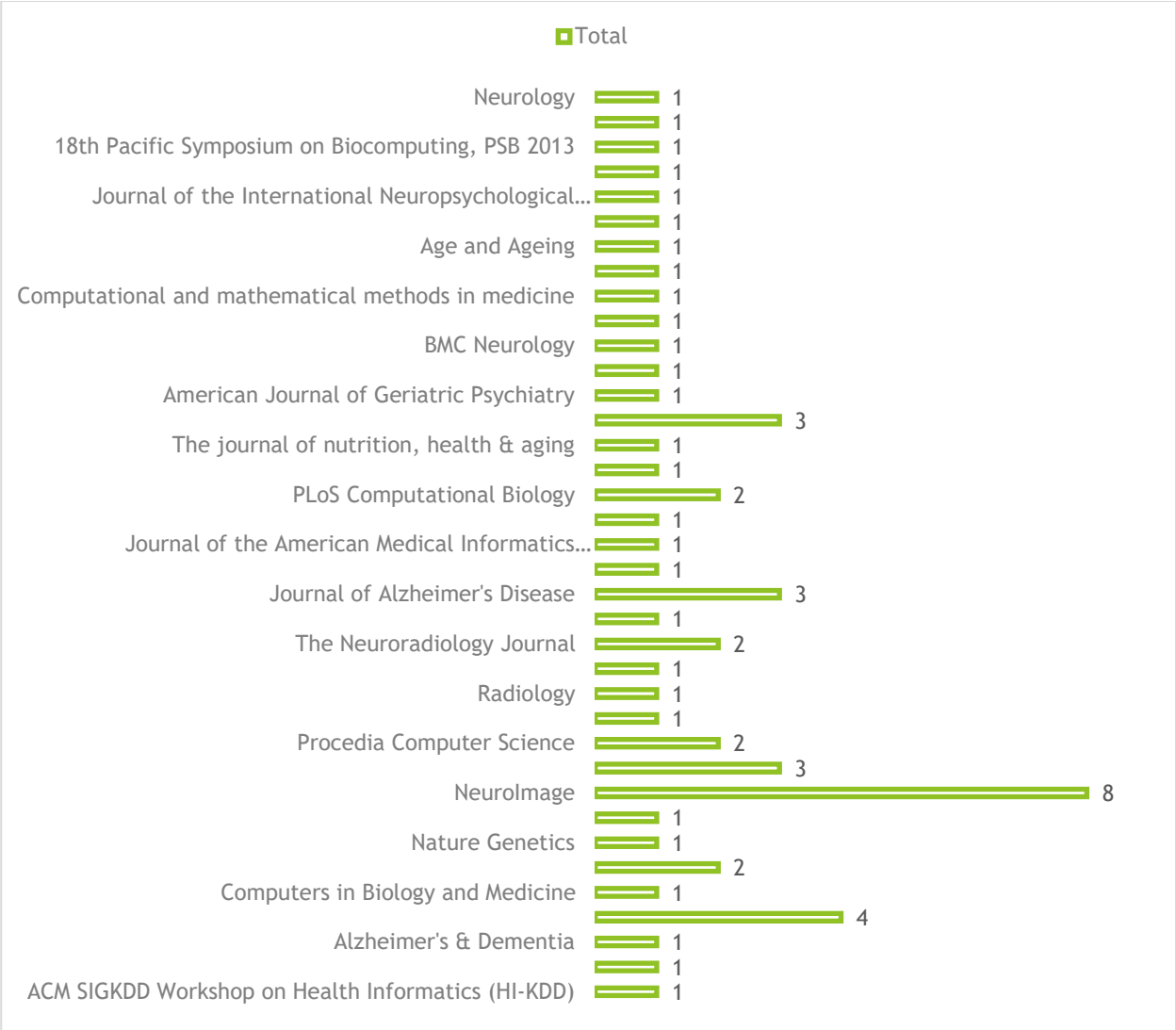


*Figure 4 – Total number of scientific articles found and to be included in the database, per its journal/publication.*

## 3.2. **Database Development**

After the identification of the scientific publications that suits the purpose of this work, the mapping of the methods, datasets and other characteristics of predictive modelling were extracted from them and mapped in a database. That database has a tremendous growth potential due to the increasing importance of AD as well as because there could be many other articles that were not found on the search stage process of this work. However, the database was developed using a Microsoft Excel Spreadsheet due to the low number of publications found. The content mapping process, the methodologies followed for data extraction, and the reference data used for the mapping is explained in this section.

### 3.2.1. **Data Extraction Model**

Following some classical approaches (Krippendorff, 1980), a quantitative systematic content analysis was followed for the articles found. After the content analysis and review, of all the articles identified as relevant, they were categorised.

For the categorisation, the model CRISP-DM was chosen to be followed (Figure 5), so that the extraction of the relevant data to be stored in the supporting database of the visualisation platform would be possible.

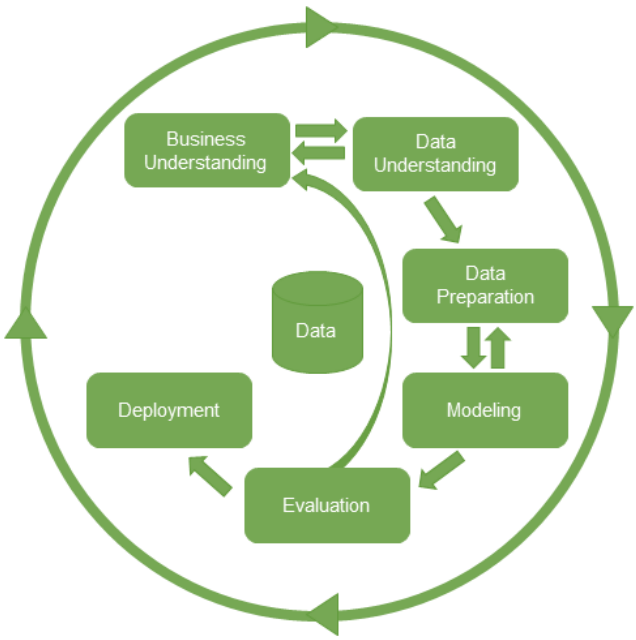


*Figure 5 - CRISP-DM methodology*

Other knowledge discovery process, "The KDD process for extracting useful knowledge from volumes of data" is presented in Figure 6. This process, presented in the first time in 1996, sets the ground for what came next in the era of data science, in the form of a framework, where possible knowledge could be obtained by retrieving and working on the significant amount of data from a given database, nowadays referred as *data mining.*

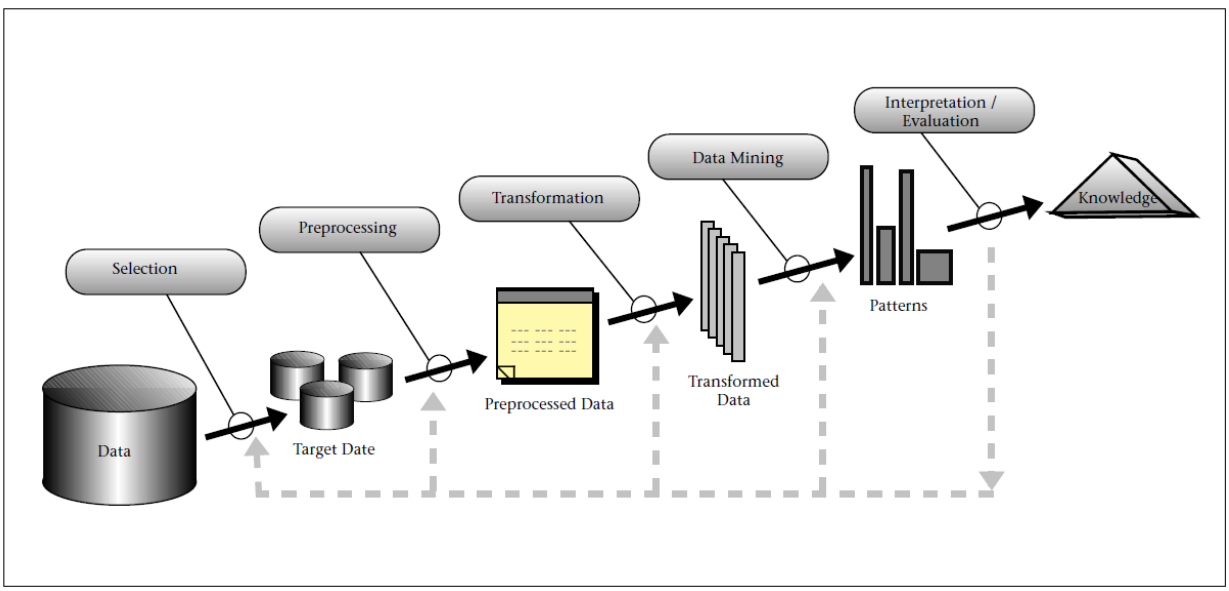


*Figure 6 - An Overview of the KDD Process (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).*

The CRISP-DM and KDD processes are very similar (Ipp, Azevedo, & Santos, 2004). CRISP-DM has the advantage of integrating the business understanding before the actual data-driven and focused phases, which serves better the purpose of this work, in the sense that it helps to contextualise the objective of each scientific paper so that it can also be mapped appropriately in the database and the visualisation framework.

Blending the two processes for the objective of this work, the process pictured in Figure 7 was followed, which is a simple upgrade to the KDD model to allow the inclusion of the *Business Needs* (under the scope of this work, it will be the main topic of each scientific paper).

Regarding the "Data Integration" step, the methods usually employed during the application of predictive analytics to the context of AD's discovery are usually the same as in the "Data Preparation" phase. Therefore, that stage will not be emphasised and was not included in the database.

Regarding the "Deployment & Monitoring" stage, it was not considered as it not under the scope of this work; this would require the verification if the method or model developed and presented in the scientific paper were eventually tested in a real medical environment.


*Figure 7 - Knowledge Discovery in Databases and Data Mining.*

### 3.2.2. **The Analogy with Business Intelligence**

In an analogy to the Business Intelligence (BI) methodologies, in order to develop the database and to design better data visualization solutions with the right attributes, measures, parameters, and so on, some initial question was posed to guide the development of the appropriate visuals in a reader's perspective. The big difference between this work and a pure Business Intelligence one it the non-existence of numerical variables in the attributes as the content is only based on categorical variables (this work is based on a qualitative problem rather than in a quantitative one). Furthermore, this analogy with BI is following the well-known Kimbal-based solution approach, where one has to come up with business problem, measures, dimensions, granularity, so on and so forth.

These questions posed to support and guide the author on the development of the database and the data visualization tool were:

1. What have been the main research objectives of the studies on Alzheimer's Disease where data science and machine learning techniques are applied?
2. What have been the most used classifiers in the modelling phase in those studies?
3. What have been the most used datasets for research purposes so far?
4. What types of data have been used in the studies carried out in the field?
5. How many articles used more than one type of data in their studies?
6. From the lists of data types and classifiers presented in Table 3, were them all identified in the articles found?
7. What are the relationships between datatype and datasets used?
8. Is there a trend regarding data pre-processing methods used in the field?
9. It there a correlation between the data types chosen and the classifier used in the studies?

10. What have been the most and the least used modelling evaluation criteria and metrics?

To answer these questions, data transformation was needed to be done in Power BI, with Power Query Editor, so that the data could fit the visuals chosen to fully demonstrate state of the art about the Alzheimer's Disease research.

As a result, the baseline for the database and the ETL developed, explained further in this report, was a list with all the articles and their characteristics, with the attributes depicted in Table 2. This list would be transformed in order to allow flexibility in the data visualization tool (explained in section 4.1.1). Note that each article can have more than on classifier, assessment method, and so on. The content of those attributes was comma separated.

*Table 2 – Fields that compose the Baseline for the DB development.*

| **List of Articles** |
|---|
| id_article |
| article_code |
| title |
| Publication |
| publisher |
| year |
| Main Research Objective |
| Data Types |
| Datasets source |
| Data Preparation |
| Classifiers |
| Assessment |

### 3.2.3. **Content Mapping**

Taking into consideration all the steps presented in the *Knowledge Discovery in Databases and Data Mining* process, and using the work done by Carreiro et al.(2015) and by R. Zhang et al.(2017), an extensive list of the leading research objectives, usual datasets, data types and formats, biomarkers and data mining methods applied to the Alzheimer's disease discovery was created.

The summary of that information is presented in Table 3**,** which aimed to be a guideline for more precise mapping of the content found in the articles used for this part of the work but does not intend to be a complete list of what has been done in the area. Moreover, what one can expect to find in the scientific literature. As it will be seen further in the report, in the results section, more was found besides what is listed in Table 3**.** It is worth remembering that this work is focused on the predictive analytics sub area of data mining (i.e. supervised learning); the descriptive methods used for classification, such as Self-Organizing Maps (SOM), were also found in studies together with the former. This is why in the database developed for this work one can find unsupervised classification models as well, in spite of being out of the scope of this work.

*Table 3 – Data types, dataset sources and classifiers used in AD research.*

| Step | Topic | SubTopic/Comments |
|---|---|---|
| **Main Research Objective** | Diagnosing AD or MCI | |
| | Predicting MCI to AD conversion | |
| | Stratifying risks for AD | |
| | Mining the literature for knowledge discovery | |
| | Predicting AD progression | |
| | Describing clinical care for persons with AD[1] | |
| | Understanding the relationship between cognition and AD | |
| **Data Pre-processing** | Filter Methods | Correlation |
| | | Minimum Redundancy Maximum Relevance (mRMR) |
| | Wrapper Methods | PCA |
| | Dealing with Missing Values | multivariable regression models |
| | | expectation-maximization |
| | | iteratively singular value decomposition |
| | Discretisation | For data simplification, while minimising information loss |
| **Data Types/sources** | Transcriptomics | |
| | Genomics | Apolipoprotein E (APOE) |
| | Proteomics and Metabolomics | Cerebrospinal fluid (CSF) |
| | Neuroimaging Data | computed tomography (CT) |
| | | transcranial sonography (TCS) |
| | | structural and functional magnetic resonance imaging (s/fMRI) |
| | | including diffusion-weighted imaging (DTI) |
| | | arterial spin labelling (ASL) perfusion; |
| | | Positron Emitting Tomography (PET) |
| | | Single-photon emission computed tomography (SPECT) |
| | Neuropsychological Data | Mini mental state exam (MMSE) |
| | | Memory impairment screen (MIS) |
| | | General practitioner assessment of cognition (GPCOG) |
| | | California verbal learning test (CVLT) |
| | | Trail making tests (TMT) |
| | | AD assessment scale-cognitive subscale (ADAS-Cog) |
| | | Iowa gambling test |
| | | Clock drawing test (CDT) |
| | | Montreal cognitive assessment (MOCA) |
| | | Logical memory (LM) |
| | | Verbal paired-associate learning (VPAL) |

[1] Although identified as a main research topic by the authors of the article (R. Zhang et al., 2017), this topic was not included in the database of articles as it is out of scope of this work.

| | | |
|---|---|---|
| | | Digit span (DS) |
| | | Wechsler memory scale-logical memory |
| | | Auditory verbal learning test (AVLT) |
| | | Hayling test |
| | | Edinburgh cognitive and behavioural ALS screen (ECAS) |
| | | Addenbrooke's cognitive examination-revised (ACE-R) |
| | | Cambridge Behavioural Inventory |
| | Neurophysiological Data | Transcranial magnetic stimulation (TMS) |
| | | electroencephalography (EEG) |
| **Datasets** | Alzheimer's Disease Neuroimaging Initiative | Demographic, neuroimaging data (MRI and PET), genetics, clinical data, (cerebrospinal fluid (CSF), cognitive tests and others), and temporal information. |
| | AddNeuroMed study | Demographic, neuroimaging |
| | Mayo Clinic Study of Aging (MCSA) | Demographic, image, genetics, clinical data |
| | ZARAgoza DEMentia DEPression (ZARADEMP) study | Demographic, cognitive measures, physical measures, questionnaires |
| | Clinical Practice Research Datalink | Medical resources utilization (consultation, specialty referral, length of hospitalization) |
| | Mendelian inheritance in man (OMIM) | Genotype-phenotype associations related to human diseases |
| | Genetic association database (GAD) | Genotype-phenotype associations related to human diseases |
| | Human Protein Reference Database (HPRD) | physical molecular interactions |
| | Biological general repository for interaction datasets (BioGRID) | physical molecular interactions |
| | IntAct | physical molecular interactions |
| | Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) | tool integrating several different databases |
| | iRefIndex | tool integrating several different databases |
| | Bio Warehouse | tool integrating several different databases |
| **Classification & Modelling** | Aggregate Measure Using Euclidean Distance | |
| | Bayesian Gaussian Process Logistic Regression (GP-LR) | |
| | Bayesian Network | |
| | Bayesian Outcome Prediction With Ensemble Learning (BOPEL) | |
| | Correlation Analysis | |
| | Cox Proportionality Hazard Model | |
| | Deep Convolutional Neural Networks | |
| | Deep Learning | |

| | |
|---|---|
| | Elastic Net Logistic Regression Logistic |
| | Linear Regression |
| | Logistic Regression |
| | Machine Learning Based Disease State Index (DSI) Method |
| | Multi-Modal |
| | Multi-Modal Multi-Task (M3T) Learning |
| | Multivariate Model |
| | Naïve Bayes |
| | Principle Component Analysis (PCA). |
| | Random Forest Classification |
| | Regularised Logistic Regression (RLR) |
| | Support Vector Machine |
| | Support Vector Machine With Particle Swarm Optimisation (SVM-PSO) |
| **Data Modelling Assessment** | Mean Square Error |
| | R2 |
| | AUC |
| | Mean Absolute Difference |
| | ROC |
| | Confidence measure |
| | Brier score |

### 3.3. **Data Visualization Framework Development**

As was already mentioned in the Literature Review section, the interest in data visualisation has been growing tremendously throughout many industries. Alongside with it, a subsector of the information system's industry has grown with the focus of bringing to life software solutions capable of drawing the patterns, trends, scales and correlations that can be drawn from the data and information available to the end user. Those analyses could be, somehow, hidden due to the underlying difficulty of a fast AND conventional data analytics approach, based on extensive unreadable tables. Furthermore, there are, as today, one branch of the data visualization that is focused on the retrieval and visualization of the scholarly data, as mentioned in the same Literature Review section of this report; nowadays, users have plenty of options to work that "science of science" data towards a more visual analytics approach.

#### 3.3.1. **Preliminary Assessment**

The goal was to create a framework for the development of an interactive data visualisation platform. It should be able to express, in a visual, natural, clear and user-friendly way, what is state of the art about the usage of predictive analytics methods to the discovery of Alzheimer's disease, with the goal of helping researchers, academics, students and institution's decision makers. The software tools and the visual

development frameworks considered are presented in Table 4 and Table 5, which were the ones offering the best baseline characteristics for achieving that main objective of the work.

The tools available can be divided into two main categories (Liu et al., 2018): tools without the programming language and tools based on a programming language.

### ***Software Without the need for Programming***

*Table 4 – List of Tools Without Programming Language.*

| **Software** | **Description** | **(Dis)advantages** |
|---|---|---|
| **Tableau** | Commercial desktop software for business intelligence and analytics and one of the market leaders in the area. Offers strong data visualisation capabilities enabling users to create intelligent business reports, dashboards and "stories" through its storyboarding feature. This software can be linked to the data files on both local and server it has an extensive data source connection. It has already a strong and big community of users. | Easy to use and to visualise the data. Very customizable and flexible and users only have to drag-and-drop the selected object, can combine multiple different types of datasets into the dashboards. It does not allow integration of third-party custom visualisations (e.g. D3.js). Present a broad range of visualisation options and a user-friendly interface for non-technical users to quickly and easily create customised dashboards, which is a significant asset for the objective of this work. |
| **Microsoft Power BI** | Commercial BI software which offers a state-of-the-art self-service business intelligence capability to the end-user, where they can create reports and dashboards by themselves. It has already a strong and big community of users. | It is a very user-friendly data visualisation tool, with drag-and-drop functionality along with access to a vast selection of visually appealing data visualisation applications. A user-friendly interface permits non-technical users to quickly and easily create customised dashboards and uses the familiar Excel functions, so experienced users in MS Office will have a smooth adaptation to this visualisation tool. |
| **ICharts** | Commercial web-based business Intelligence reporting and analytics tools for *Netsuite*, Google Cloud platforms and other databases, being able to combine enterprise data from different sources | Real-time business intelligence tool, avoiding the need for a manual update; databases from different sources are linked so that it can automatically update the content. Flexible but there are no unconventional charts or graphs. |
| **Infogram** | Web-based application with many resources and options for the creation | Friendly user experience, vast public chart type library to get inspiration. It also |

| | | |
|---|---|---|
| | of visual content, such as charts, maps, graphics, and dashboards. Infographics look to be the main creative focus of the tool. | provides real-time data processing and supports multi-terminal displays. The disadvantage is that the uploaded data in the tool's online database is public (unless the user chooses for a paid membership plan). |
| **RAW Graphs** | It is an open source web-based data visualisation framework built for democratising the visual representation of complex data, having a very responsive UI. It can be seen as a link between spreadsheet applications and vector graphics editors. | Data processing only using the local web browser which ensures data privacy. Offers a good range of charts and some unconventional ones, as well as it allows the creation of custom vector-based visualisations on top of the D3.js easy to use. Good export options and user-friendly but unsupported web application publishing of the customised charts. |
| **Visualise Free** | Free and cloud-hosted web-based business intelligence application with data preparation, visualisation and analytics capabilities. It is a derivative of a commercial platform for dashboard, reporting and data mashup developed by the same company. | Very responsive web application. Users can only upload a rather small data file. It is easy to use and suitable for users to visualise a small quantity of data either in common or in exquisite charts, graphs and maps. Users can easily visualise their data with multiple beautiful charts by dragging and dropping the data into the correct layout. Free visual analytics is provided, and then users can compile a detailed analysis of the uploaded data (which will always be private). The generated charts can be shared by moving them into the shared folder or downloading them in the chosen format. |

Other commercial and privately owned BI software were not considered in the preliminary analysis such as Sisense, IBM Cognos Analytics, SAP Business Objects, Domo, Oracle BI, as these only have standard enterprise-focused and business-driven visuals, whereas Microsoft and Tableau seem to be democratising data visualization and Business Intelligence with a wide range of their beautiful and artistic-like new generation visuals (Hagerty, Sallam, & Richardson, 2011).

*Tools/Frameworks that Require Programming*

For the frameworks described in this section, the main disadvantage is the need for software/web development skills and the time it requires, in comparison with the tools described above, is the most significant advantages the flexibility it gives to the visuals developer and the way it can be customized to the input data and the end user.

*Table 5 – Tools/Frameworks that Require Programming.*

| *Framework* | Characteristics |
|---|---|
| D3.js | It is a JavaScript library for manipulating documents based on data, helping to bring data to life using HTML, SVG, and CSS and using the full capabilities of modern browsers. Without being dependent solely to a proprietary framework, combining powerful visualisation components and a data-driven approach to DOM manipulation. The interactive chart and graphs it generates are all in .svg format. It is requested to run with its official document library for function invocation. |
| Chart.js | It is a program of the open source JavaScript graphics library; it uses canvas on HTML5, it can visualise the data into several common chart types by invoking the script language and its official chart library. |
| Fusioncharts | Commercial JavaScript library suite that combines JavaScript and ActionScript3.0. It can run on multi-devices, browsers, and platforms. *Fusioncharts* has many chart types and maps, and it supports the.xml and .json files processing. The exported charts can be done to .jpg, .png and .pdf file types. The generated interactive charts It can be embedded in user's applications with several wrappers (e.g. JSP charts, PHP charts, jQuery). |
| Flot Charts | It is focused on simplicity basic (charts-bar, line, point, and segment). It is an extension of the jQuery library that supports HTML5 charts which are combined canvas and VML. This library separates the functional logic from HTML structure and uses DOM (Document Object Model) element to complete plotting. |
| Zingchart | It integrates Angular, React, jQuery, PHP, Ember, and Backbone in its declarative, efficient, and simple JavaScript library. Besides the offered standard charts and It also offers integrated chart arrangement capabilities and has the handy drill-down function. |

Other tools and platforms that based on programming languages were put aside (e.g. Gephi, GGPlot2, JPGRAPH) as they are more useful for visual analytics rather than data visualisation and thus, not suitable for the data visualisation solution to be developed.

### 3.3.2. **Tool Selection**

For the narrowing down of the found options to the one used in this work, a more in-depth assessment was carried out, putting an effort on describing the advantages and disadvantages of two of them: Tableau, Microsoft Power BI and the JavaScript's library and data visualisation's framework D3.js. The selection followed four last main criteria: connectivity to data sources, diversity of visuals, flexibility for further adaptation and integration with third-party applications and development effort.

*Table 6- Criteria assessment for the Tableau, Power BI and D3.js.*

| Criteria | Assessment |
|---|---|
| *Connectivity to Data Sources* | As the database with all the data extracted from the articles was built in Microsoft Excel, not being required to be connected to another internal or external databases, data warehouses and other data sources, the requirement is not that restrictive. Therefore, all the three cases passed in these criteria |
| *Diversity of Visuals* | Visualisations are one of the core focus of this part of the work. Therefore, although Tableau provides to the end user a very intuitive interface, Power BI has an advantage over Tableau due to the possibility of importing custom visuals in the tool from its app store, and some of them are much more appealing and suitable for the specific objective of the visualisation framework of this work. Furthermore, D3.js is, perhaps, the most exciting option for this work under this topic, having quite interesting visuals and chart/graph design. |
| *Flexibility* | As already mentioned in the previous topic, Power BI is quite flexible as it allows the import of different and custom visuals in the tool from its app store, including the integration of R programming language scripts for data visualisation and, actually, D3.js data visualisation scripts. Furthermore, it is the only one of these three data visualisation and analytics apps that have extensive R and big data-related integrations, ensuring a good scalability prospect for further works on top of this one. |
| *Development Effort* | The time to develop a  using Javascript and D3.js based webpage from scratch is much more than using the drag-and-drop user-friendly environment either of Tableau and Power BI, in particular for the ones not experienced in that particular programming language. |

Not only for the sake of time efficiency, to not compromise the delivery, but also from the cost(time)-benefit standpoint, the option of using one of the commercial software's stands out by being the most appropriate one. Thus, overall, Power BI emerges as the most suitable solution among the three selected tools.

### 3.3.3. **Design, UI and UX**

The interactive visualisation platform tries to integrate the gathered knowledge and inspiration from experienced designers, data and information visualization expert's publications like Healy (2018), McCandless (2012), Munzner & Maguire (2014) and Yau (2011), but taking always into consideration the limitations of Power BI itself (for instance, it does not support version 4 of the D3.js library).

The process of creating a story with the data at hands, aiming for the storytelling mindset (Healy, 2018; McCandless, 2012; Yau, 2011), is now more possible and that was the idea behind the development of the interactive visualization tool. With the rise of business intelligence visualization software, storytelling creation is way simpler and faster and, often, much more appealing than before.  From a user and reader perspective, either from the scientific environment already, thus already knowledgeable about the subject, or completely outside from the field, one will want to (1) understand what the main objective of the work was, (2) what data and the data analytics methods were used and lastly (3) what were the main

conclusions. This led to the conceptual model depicted in Figure 8, in line to what was also presented about the CRISP and KDD models in section 3.2.1. One should note that the block named “Main Research Conclusion” are not under the scope of this dissertation”.

The data visualisation’s interactivity was deeply considered. Having the possibility of working with D3.js framework and development library inside Power BI’s business intelligence tool was a key factor for using this tool, which is a flexibility enhancer to address that need. Furthermore, different and innovative custom visuals (publicly available at Microsoft App store for Power BI product) were experimented to reach to the final ones that are included in the final solution. At the same time, as D3.js is based on an open framework, it can be enhanced and further developed in the future (discussed in the section Limitations and Future Work).

To implement such a visualization framework, some consideration has been made. For instance, it was not possible to use a Tree Map as all the nodes are interconnected with each other, meaning that a particular article can use both a Neural Network and Logistic Regression which would result in a duplication of tree leaves using this visualisation design concept. The same happens with datasets and data types as most of the articles used more than one.

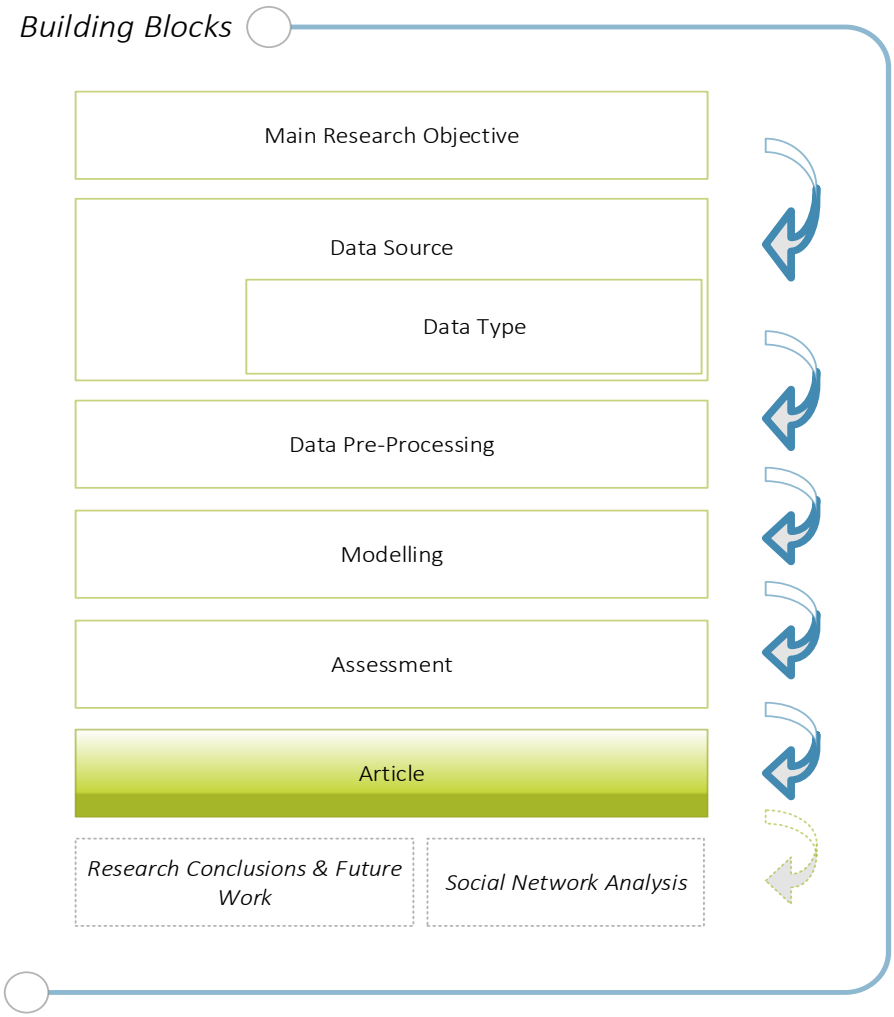


*Figure 8 – Building blocks of the Data Visualization Framework Conceptual Model.*

The implementation of this framework required the correct mindset concerning possible future upgrades. Therefore, a *connected blocks* structure was developed which was thought to be the better solution to depict the database structure, content and its purpose. Furthermore, as depicted in Figure 8, it can be seen as a flow of activities, meaning that the researcher, to carry out his/her study, would need to follow a path, starting by defining its research main objective (or question), deciding on what data type must use, which dataset can provide that selected type of data, select the appropriate classifier(s) to train the best possible model and then assess it by using the right metrics for its model.

In a next phase, not considered in this dissertation but summarised in the section Limitations and Future Work, the Main Research Conclusions and the other metadata, i.e. authors, institutions, and so on, would be integrated, being the latter used to integrate a Social Network Analysis tool, such as the ones described in the Literature Review section, in order to provide a complete, multipurpose holistic view of state of the art research in the field.

# 4. RESULTS & DISCUSSION

After the database development phase, with the article content mapping finished, it was possible to work with the final database. The database and visualisation tool produced from the followed process explained in the *Methodology* section is found in this chapter.

## 4.1. **Scientific Publication Database**

### 4.1.1. **Transactional Relational Database**

A simple Extract-Transform-Load (ETL) process was developed inside Power BI after the import of the list created in Excel, not only to follow the good practices of database management but also to allow more flexibility in the data visualization tool. A star schema relational database was developed from the list created in MS Excel (Table 2), after data transformation. In order to speed up the queries later on, it was decided to avoid repetition of entries in the table *Fact_Article*, as each one can have multiple entries of classifiers, data types and so on. Therefore, this was a needed step due to the nature of the data, the objective and because it would allow further developments in the future – it is much more flexible for the integration of other parts and content from the scientific articles, apart from the ones considered in this work.


*Figure 9 – Star Schema shaped Transactional Relational Database.*

The relational database has one Fact table called *Fact_Article*, and six dimensions. All the non-*id* variables are categorical and have the data type set as [text] (apart from the year, which was set as [int], as Power

BI changes the Data Year to integer). Due to the nature of the problem, no measures were needed in the Fact table at this stage. Due to the low volume of data for this database, it was not needed to develop a data warehouse.

It was created three hierarchies with the tables in which subtypes exist, namely in table *Dim_Classifier*, *Dim_DataSource* and *Dim_DataType* (Figure 28, in the Annex).

One important aspect about the content of the table is that some fields have, on purpose, the content "N/A". It is not a *null* value but rather a flag indicating that either it was not possible to access the full text of a particular study or it was not clear enough which method, technique, data types, and so on, if any, were applied at that particular stage of the KDD process.

### 4.1.2. **Data Analysis**

In this section one can see the main data analysis of the dataset used for the development of the visualization tool. As a summary, which shows the maximum, average and minimum number of items of each KDD step for the articles included in the database.

*Table 7 –KDD statistics for the articles included in the database.*

| | **Datasets** | **Classifiers** | **Assessment Methods** | **Data Pre-Processing Methods** | **Data Types** | **MRO** |
|---|---|---|---|---|---|---|
| *Max* | 2.0 | 5.0 | 5.0 | 10.0 | 14.0 | 2.0 |
| *Avg* | 1.1 | 1.6 | 3.4 | 2.1 | 3.5 | 1.0 |
| *Min* | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |

#### **4.1.2.1. Main Research Objective**

The Main Research Objective is the starting point of the research. Therefore, it was the first step to be analysed. To apply predictive analytics methods to the Alzheimer's Disease field is usually with the objective of either "diagnosing AD or MCI" or "predicting MCI to AD conversion". These were the two significant objectives found in the articles gathered and available in the database. The other four types of main research objectives are evenly distributed among them Figure 10. More tables and charts can be found in the Annex.

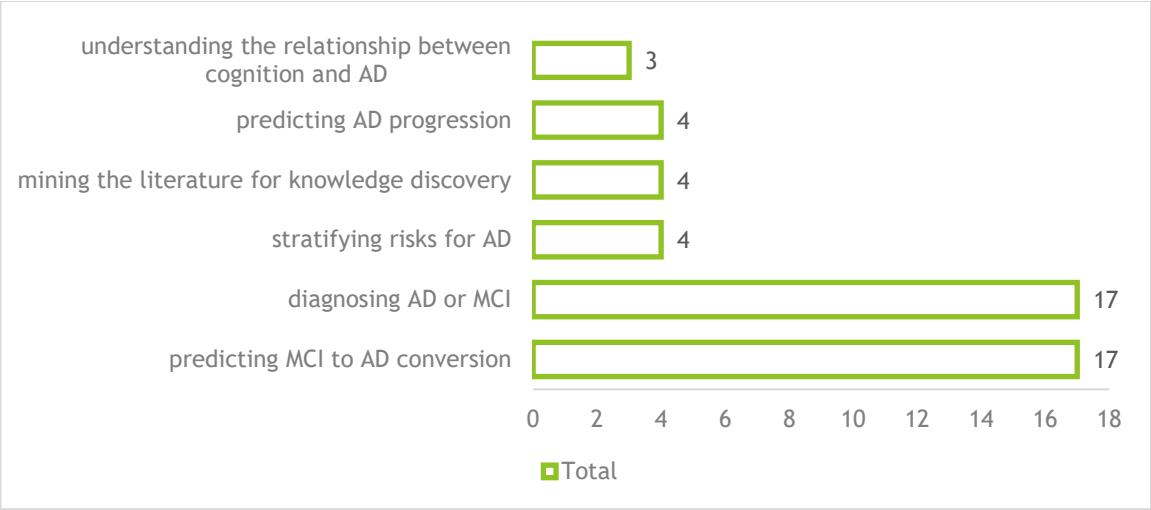


*Figure 10 – Total number and distribution of the Main Research Objectives of the research articles.*

#### 4.1.2.2. Data Types

The data types used in the research articles found, and included in the database, are the ones in Figure 11. Neuroimaging is precisely the data type most used, followed by the category of neuropsychological data, the latter including a wide set of data gathering methods usually done by psychological tests.

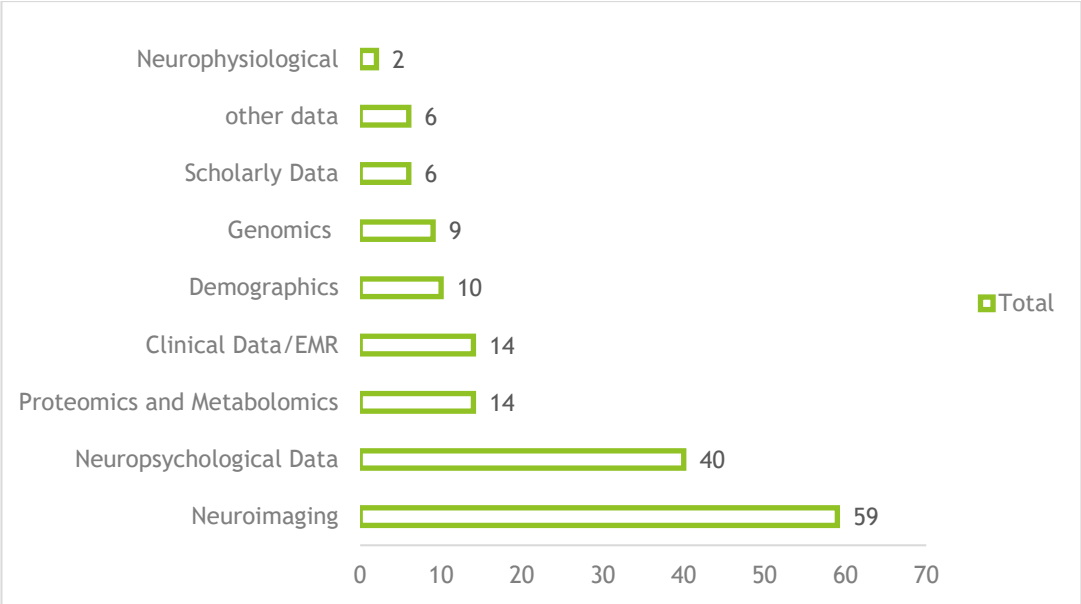


*Figure 11 – Total number of Data Types used in the research studies.*

The data type found in the articles were not necessarily what was presented in Figure 11, i.e., those are the parent category of the actual data type used. In some research articles, the specific data type used was found and, in that case, it was categorized according to Table 3. This categorisation had two main objectives: increase specificity to the data mapping from the articles to the database and to make the visualisation easier and cleaner later on. In Table 8 one can see the full list of data types and "subtypes"; while the *data_type* can be either the more low-level description of the data type used in some particular study or the name of the category, the *data_subtype* is the actual data type used whenever it was clearly mentioned in the article.

*Table 8 – Data Types and "Subtypes" list.*

| data_type | data_subtype | Count of data_subtype |
|---|---|---|
| **Genomics** | | **9** |
| | ApoE | 7 |
| | gene expression | 1 |
| | SNP | 1 |
| **Neuroimaging** | | **42** |
| | MRI | 28 |
| | MRS | 2 |
| | NB | 1 |
| | PET | 9 |
| | SPECT | 1 |
| | T2/proton density magnetic resonance | 1 |

| | | |
|---|---|---|
| **Neurophysiological** | | **2** |
| | EEG | 2 |
| **Neuropsychological Data** | | **40** |
| | ADAS-Cog | 7 |
| | Auditory–Verbal Learning Test | 1 |
| | AVLT | 1 |
| | Boston naming | 1 |
| | CDR | 6 |
| | CVLT | 1 |
| | GDS | 2 |
| | LM | 1 |
| | MMSE | 12 |
| | MOCA | 1 |
| | NIQ | 1 |
| | NPI-Q | 1 |
| | PsyMEM | 1 |
| | RAVLT | 2 |
| | semantic fluency word lists | 1 |
| | TMT | 1 |
| **other data** | | **6** |
| | DxConv | 1 |
| | plasma levels of vitamin E forms | 1 |
| | speech samples | 1 |
| | Trails B | 1 |
| | vascular risk factors and diseases | 1 |
| | Visual Paired Comparison | 1 |
| **Proteomics and Metabolomics** | | **14** |
| | AD-related protein interaction network | 1 |
| | CSF | 12 |
| | protein-disease association | 1 |
| **Scholarly Data** | | **6** |
| | article abstracts | 4 |
| | literature reviews | 1 |
| | scientific reports | 1 |
| **Grand Total** | | **119** |

#### 4.1.2.3. Datasets

Once the data type what the study/researchers will focus one, they will be trying to find data and datasets in the available data sources, either being AD's patients, people with MCI or healthy people, depending on the main research objective. The full list of the data sources used in the articles found can be seen in Figure 12. As it can be noted, the ADNI database is quite popular and widely used as the main source of data. In the next section, where the data visualisation platform is described, one can see the association between the datasets and the datatypes used.

The datasets were aggregated in some main categories, namely “Local/National Alzheimer’s Disease Research Center” and “Other Sources” so that the less usual data sources can be seen separately, which will also help on cleaning the visualisation canvas later on. The list of those called *dataset_sub_source* can be seen in Table 9.

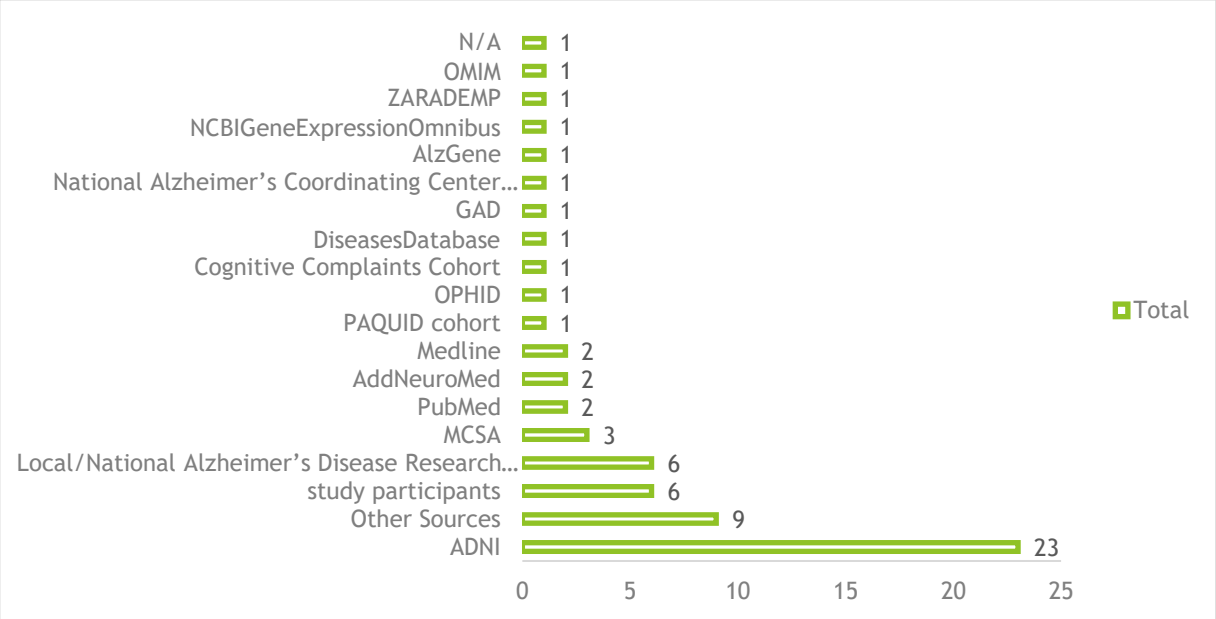


*Figure 12 – List of the data sources used by the researchers in their studies.*

*Table 9 – List of other dataset sources.*

| dataset_source | dataset_subsource |
|---|---|
| **Local/National Alzheimer’s Disease Research Center** | |
| | national Alzheimer's centre |
| | National Alzheimer’s Coordinating Center (NACC) Uniform Data Sets |
| | Wisconsin Registry for Alzheimer’s Prevention |
| **Other Sources** | |
| | alzforum.org |
| | Alzheimer Disease & Frontotemporal Dementia Mutation Database |
| | Diseases Database |
| | Gene Ontology |
| | Gensat Brain Atlas |
| | Ginkgo Evaluation of Memory study |
| | KEGG |
| | KuopioL-MCI |
| | NCBI Gene Expression Omnibus |
| | synthetic data |
| | Telemakus knowledgebase |

#### 4.1.2.1. Data Pre-Processing

The full list of pre-processing data methods is quite extensive, which is shown in the chart of Figure 26, in the Annex. There is no trend whatsoever or even a method that stands out by being the most used. Instead, a wide range of data pre-processing is used, showing the diversity of the research done in the AD discovery with the application of data science and statistical tools and methods. It is worth mentioning that when neuroimaging data is used, usually in MRIs, the same type of data processing is done (converting image data to a workable dataset which can be used for the modelling phase) and it was not mapped in the database developed for this dissertation.

#### 4.1.2.2. Classifiers

The list of classifiers, or families or classifiers, found in the articles are shown in Figure 13, and the list of specific classifiers that belong to the created classifier category (e.g. Logistic Regression category), can be found in Table 10.

Regarding the classifiers mostly used in the studies about Alzheimer's Disease, SVM and Logistic Regression are, among the most popular ones, the most used ones. Moreover, there is a wide range of classifiers in the list as well as included in the category "Other Methods", the latter being a "bucket" where models developed, for instance, specifically for the study are included.

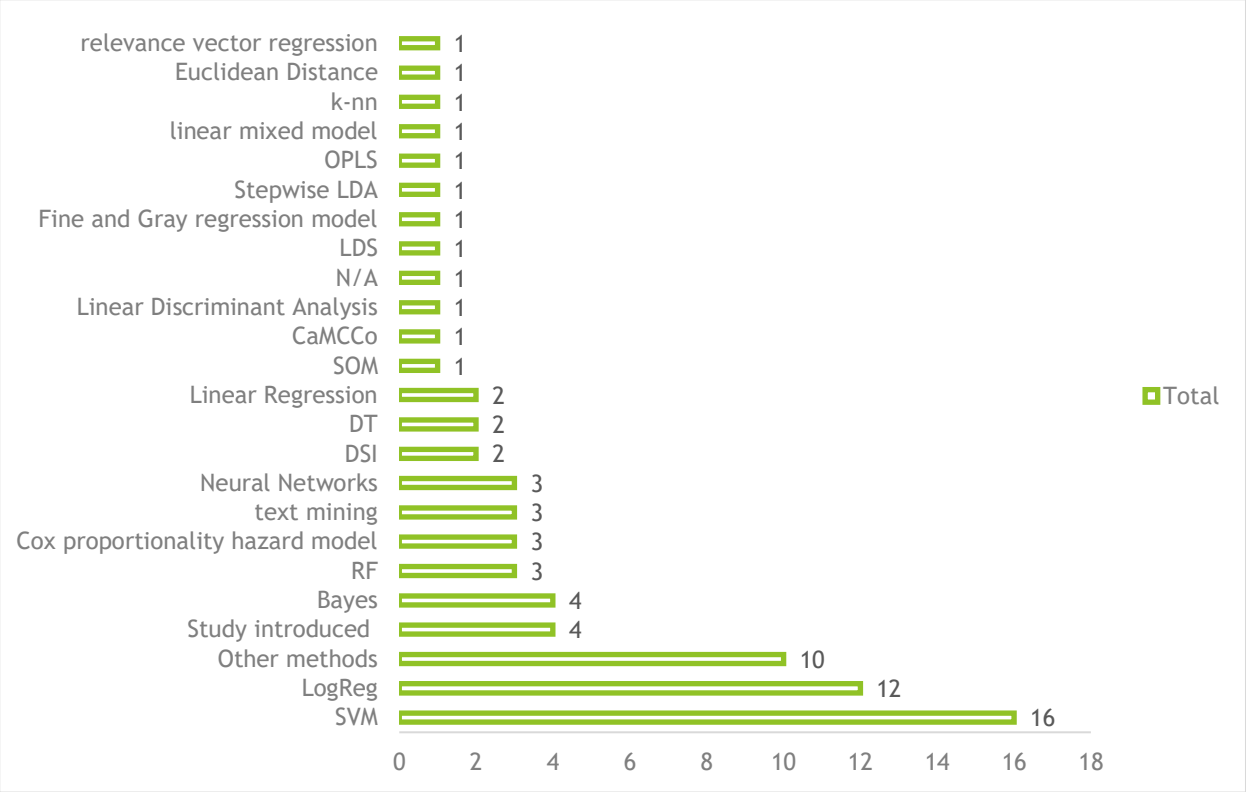


*Figure 13 – List of Classifiers and its Occurrence in the articles found.*

Regarding the variable *classifier_subtype*, it is not a subtype *per se*; it was instead a way to aggregate classifiers that belong to the same family to help further down the road on the visualisation part of the work. Due to the importance of this step of the predictive analytics process in the overall results of the research studies, this step must be readily visible and understood by the end user and reader, concerning,

for instance, the most and least used classifiers. An adequate visual with the right information is crucial for the correct understanding of the problem and to get the answer the reader is looking for. Therefore, and in line to what was already described for the datatype section above, it was opted for aggregating as much as possible the classifiers in families, at the same time that the visibility of the all the relevant information is not jeopardised.

*Table 10 – List of "classifier subtype".*

| classifier_model | classifier_subtype |
|---|---|
| **Bayes** | |
| | Bayesian Networks |
| | BOPEL |
| | Naive Bayes |
| **DT** | |
| | C4.5 Decision Tree |
| | surrogate decision tree |
| **LogReg** | |
| | bayseian gaussian process logistic regression (GP-LR) |
| | elastic net logistic regression |
| | EN-RLR |
| | regularised logistic regressions |
| | Regularized LR |
| | RLR |
| | stepwise logistic regression |
| **Neural Networks** | |
| | ANN |
| | Convolutional NN |
| | GRNN |
| **Other methods** | |
| | association analysis |
| | Chi-square |
| | Correlation analysis |
| | linear mixed model |
| | Mann-Whitney U test |
| | multi-kernel learning |
| | multi-modal disease marker |
| | Network Mining |
| | rDAm |
| | relevancy percentage |
| **Study introduced** | |
| | Hypothesisfinder |
| | iSFS |
| | Multi-model multi-task (M3T) |
| | Multivariate regression model |
| **SVM** | |
| | SVM-PSO |

#### 4.1.2.3. Assessment Methods

Regarding the assessment metrics used to compare the developed model's performance that was identified in the research papers, there was a high percentage of articles (25 articles which correspond to ~52%) that used the ones associated with the Receiver operating characteristic (ROC) curve: sensitivity, specificity and accuracy. Then there is a long tail of metrics that are also used to evaluate the quality of the models and methods proposed by those studies.

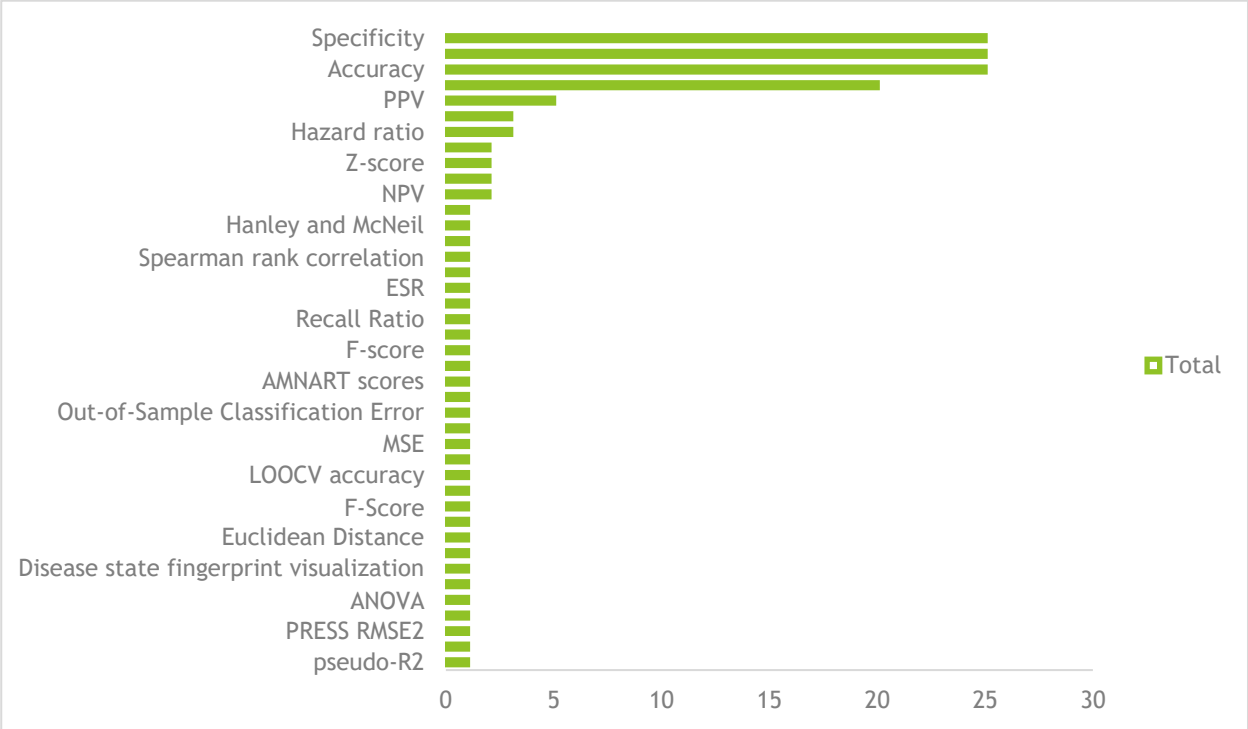


*Figure 14 – List of the assessment metrics used to evaluate the models' performance and its usage frequency over the articles found.*

## 4.2. Interactive Data Visualization Framework

The designed data visualisation framework was carried out using Microsoft Power BI Desktop, as explained in chapter Data Visualization Framework Development. It was decided to promote flexibility over a static approach. Twelve dashboards composed it. Data cleaning, homogenisation and creation of subtypes and subcategories (explained in the previous section) had to be performed, to make the data visualisation appealing and legible. Furthermore, the following topics were also considered:

1. Custom visuals for Power BI from Microsoft Marketplace were needed to be imported, namely:
    a. Sankey diagram
    b. Chord diagram
    c. Sunburst
    d. Journey Charts

2. Although in the database all the six *Main Research Objectives* exist, the following main topics were filtered out for a matter of simplicity as far as the data visualisation is concerned:
    a. “mining the literature for knowledge discovery.”, articles (
    b. “stratifying risks for AD.”
    c. “understanding the relationship between cognition and AD.”

### 4.2.1. **Dashboard Description**

Twelve dashboards were developed to fully show the interdependencies, in the form of connections, between the seven building blocks of the application of predictive models. The title of the following subsections is the same as the name of the different tabs/dashboards in the Power BI based platform.

*1 - Research Flow*

This dashboard follows the building blocks based conceptual framework depicted in Figure 8. It shows the path between the objective/question phase to the article itself, the last stage of the seven. As it was already mentioned, the Main Research Objective was filtered (for the visualisation only) to show only the three more relevant ones for this work: “predicting MCI to AD conversion”, “diagnosing AD or MCI” and “predicting AD progress”. The seven stages are represented here, and an in one part of each block highlights its dependent paths. The last block is composed by the end nodes which are the *article_code*. With that the reader can refer to the article list and get the other available metadata such as the publisher and the title. It can easily be identified, for instance, that neuroimaging and neuropsychological data are the most common data types used, and mainly for studies with the objective of “predicting MCI to AD conversion” or “diagnosing AD or MCI”.

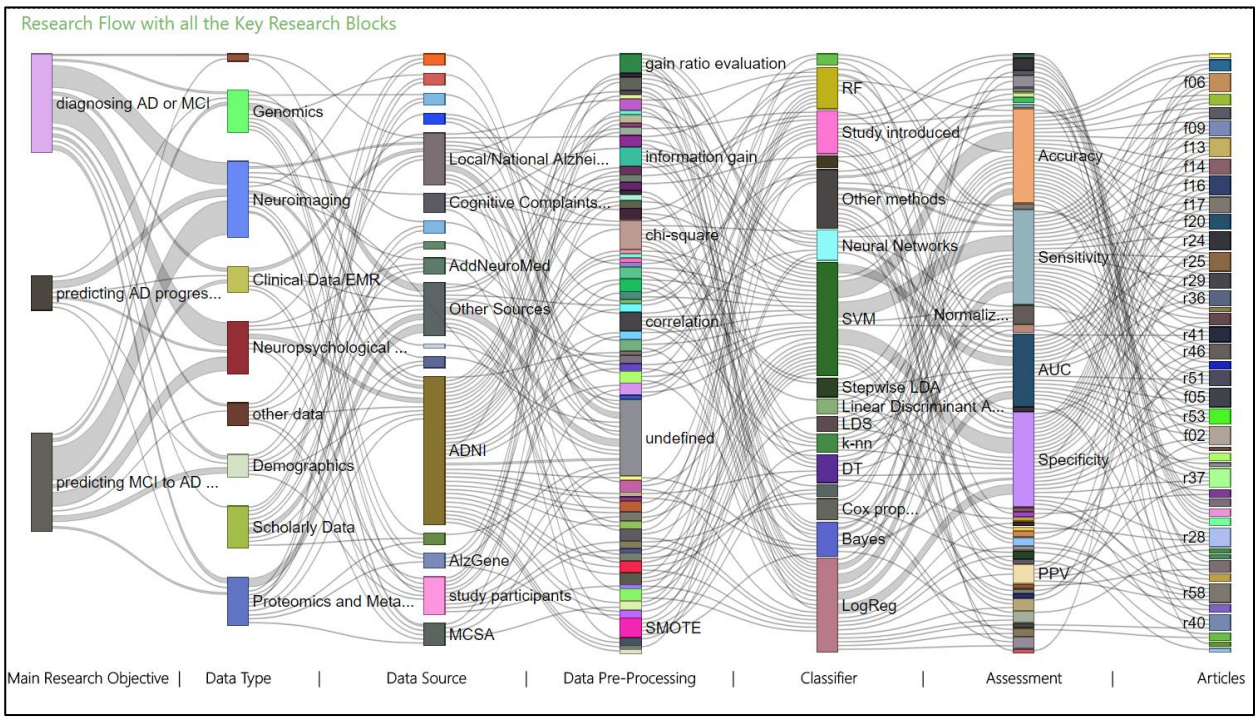


*Figure 15 – Research Flow.*

*2- Research Flow Simplified*

The data pre-processing step, explained in section, 4.1.2.1, has a wide range of values. It means that visually it is quite difficult, in the space available in Power BI window, to clearly show all the possible values without creating a bit of confusion. Furthermore, it is also more important to see the relationship between the datatype, data source and the classifier used. Therefore, this dashboard eliminates that building block compared with the previous one to address these issues.

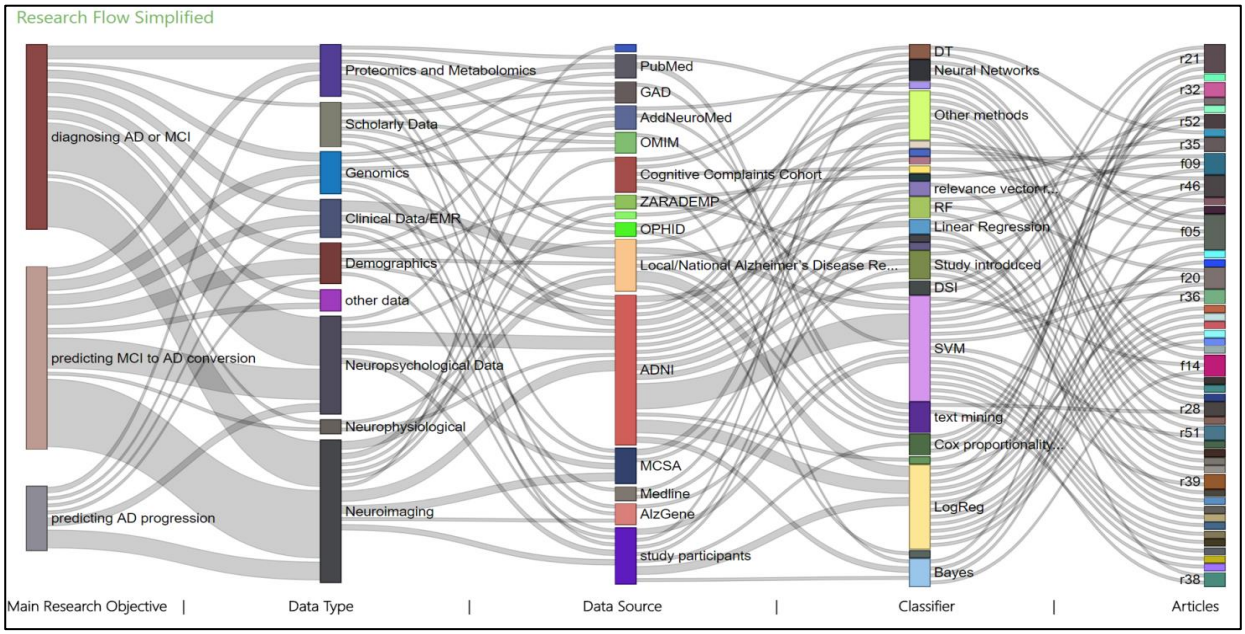


*Figure 16 – Research Flow Simplified.*

## *3- Main Research Objective, Data Types, Data Source*

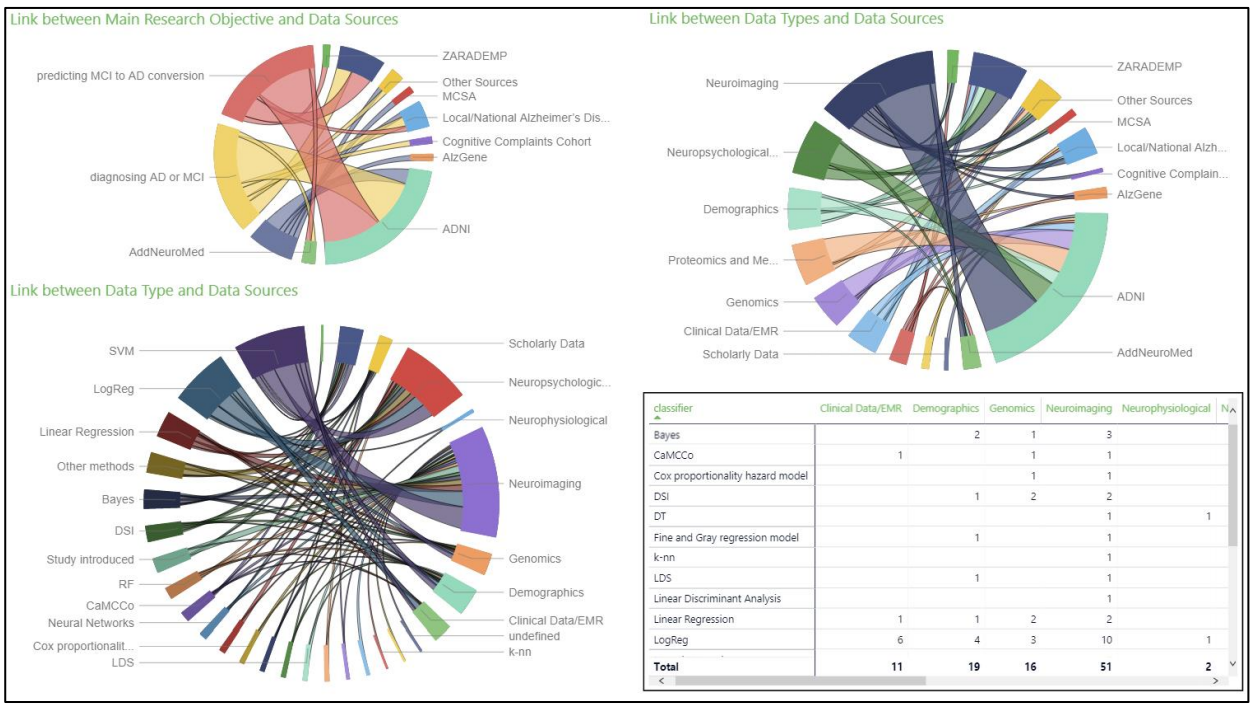


*Figure 17 - Main Research Objective, Data Types, Data Source.*

The “Chord Diagram” is a custom visual that is also available at the Microsoft Marketplace. It was used here to show the relationship, and its magnitude, between the Main Research Objective, the data type and the data source. At the same time, there is a table with classifier versus data type, so that the user can filter what it is more relevant to him/her and to quickly check the dependencies between all these building blocks. All the charts, as well as the table, are interactive and by selection one specific item it will filter the information in the other ones. Thus, one can easily see that many neuroimaging-focused studies rely on ADNI database to source their data, which demonstrates the importance of that particular database in the field of Alzheimer’s Disease diagnosis research.

*4 - Data Types*

This dashboard has the objective of showing the representation and distribution of the data types and subtypes (drill down and up is possible by clicking in one segment of the doughnut-shaped chart), shown on the left-hand side of the dashboard. It also tries to identify a trend in the usage of a particular type in the time domain, shown on the right side. As one can see, neuroimaging is the most representative category of data type used, and there is no clear trend on the usage of a particular data type as well.

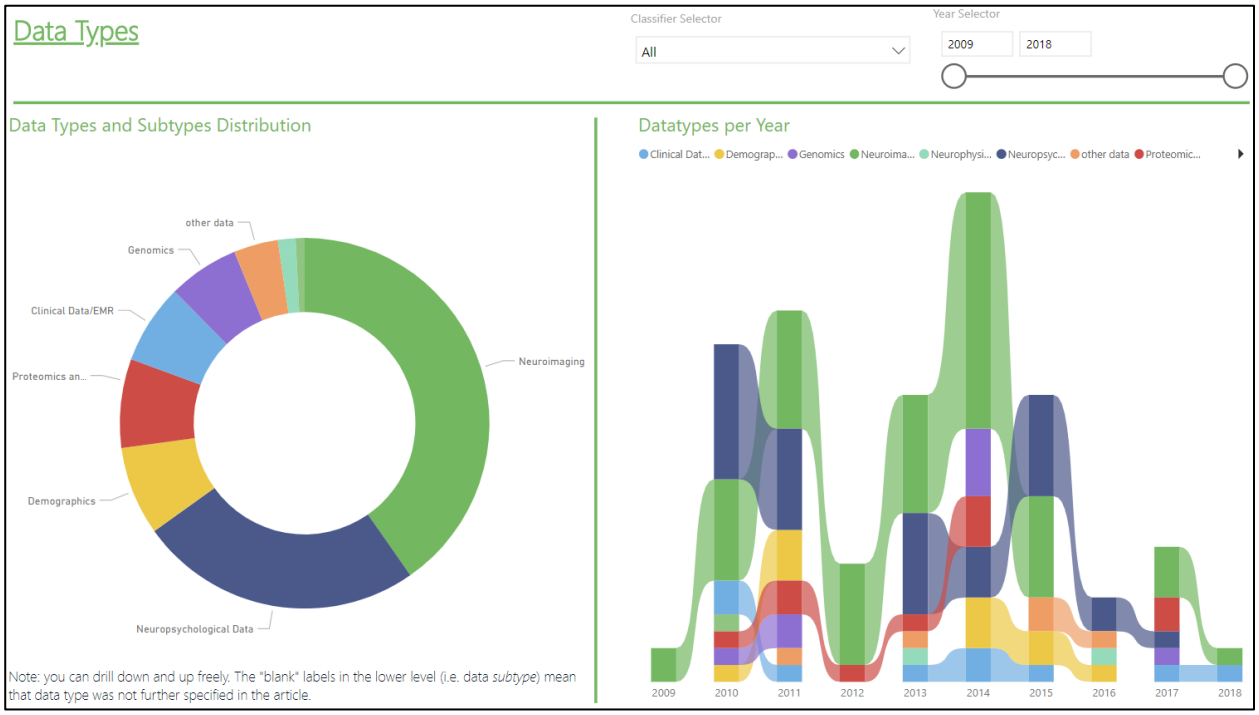


*Figure 18 - Data Types dashboard.*

*5.1- Data Sources*

This dashboard has the objective of showing the representation and distribution of the data sources/databases used by the studies (drill down and up is possible by clicking in one segment of the doughnut-shaped chart), shown on the left-hand side of the dashboard. It also tries to identify a trend in the usage of a particular source in the time domain, shown on the right side. ADNI is still widely used, and in recent years it has been one of the main datasets used by researchers.

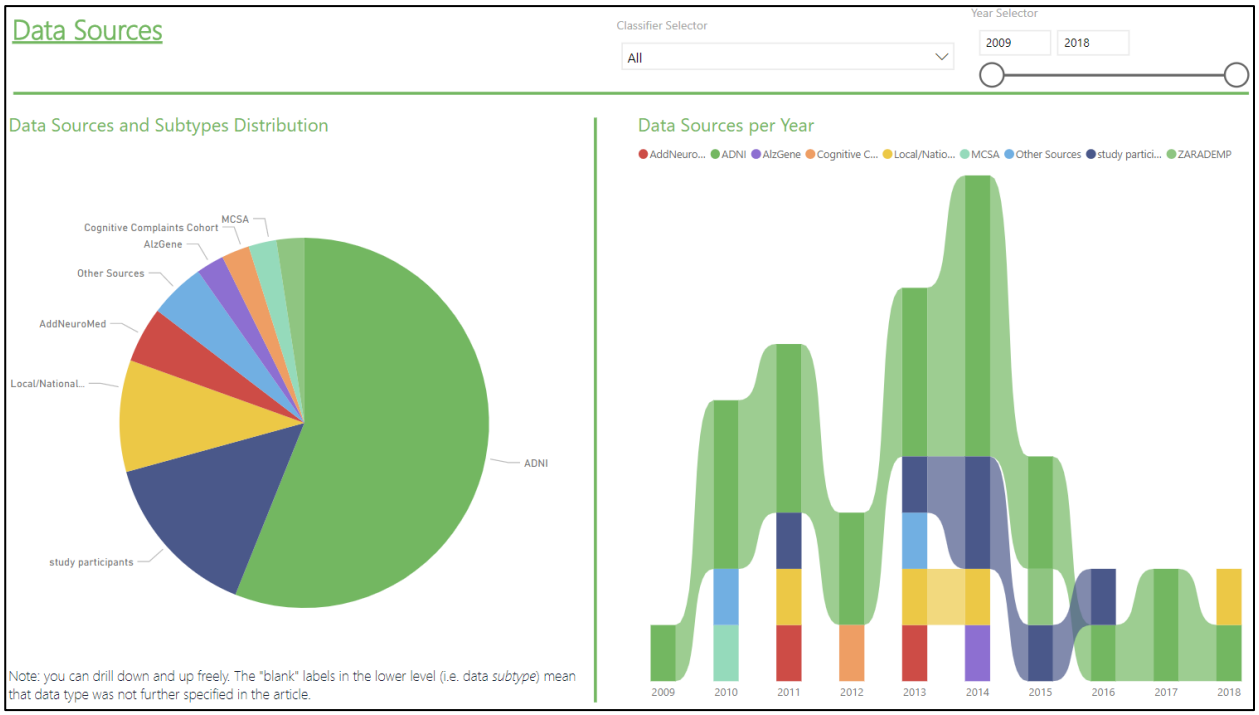


*Figure 19 - Data Sources.*

## *5.2 - Data Source & Type Subcategory*

It was decided to create this Sunburst based chart depicting the relationship between four layers: datatype, data subtype, data source and data sub-source.

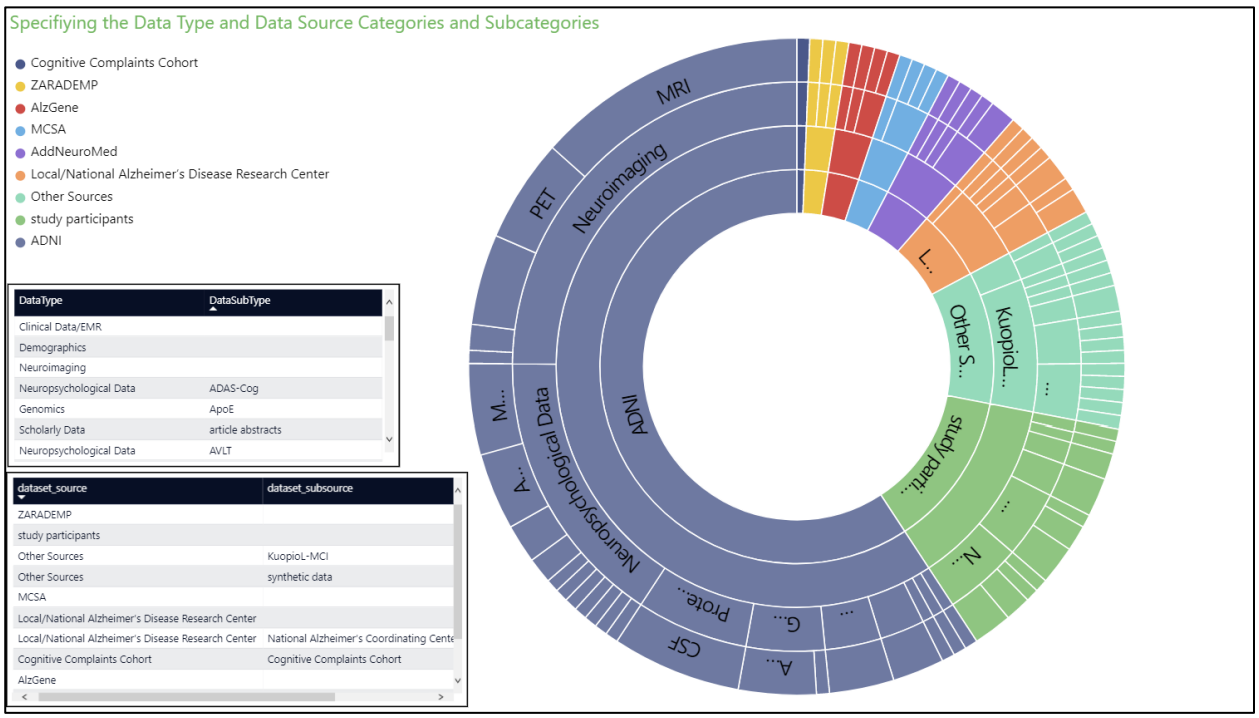


*Figure 20 - Data Source & Type Subcategory.*

It is one other way to show the data available as far as the key ingredient, data, for the predictive models is concerned. It is quite interesting to see the relationship between all that data. From this chart, one can easily see that MRI is the most common data type used, not only from the neuroimaging domain but overall.

*6 - Data Pre-Processing*

This dashboard shows the information regarding data pre-processing methods. It is possible to see that it is fairly distributed among all the methods used, despite chi-square, DARTEL normalisation, genetic algorithm, kNN, PCA and segmentation lead by a small amount of difference.


*Figure 21 – Data Pre-Processing.*

*7.1 – Classifiers*

This dashboard has the objective of showing the classifiers used in the studies as well as the evolution of their usage over time. The classifier SVM and the Logistic Regression-based models are the most popular ones in this field of science, particularly SVM that has been used consecutively over the years.

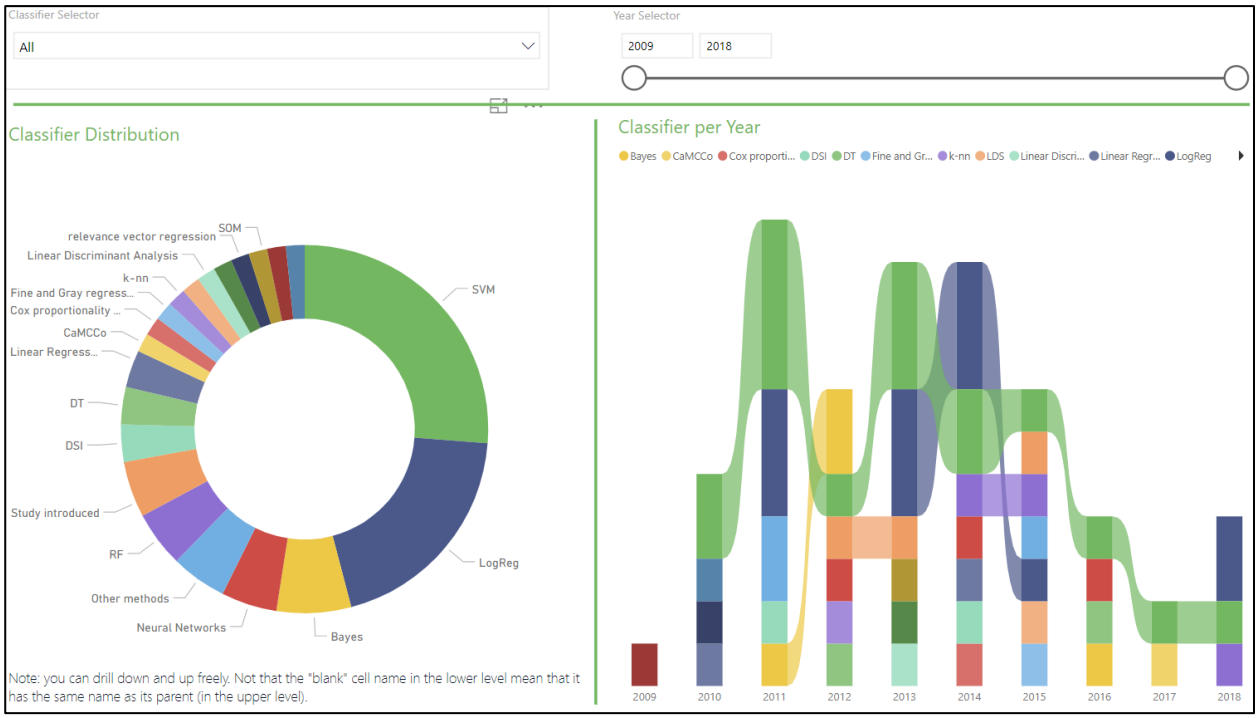


*Figure 22 – Classifiers.*

## *7.2 - Classifier Sub Category*

This dashboard has the main objective of showing clearly which models belong to the categories chosen, such as LogReg, Bayes, and so on, as that might be not clear enough for some users in the other visuals.

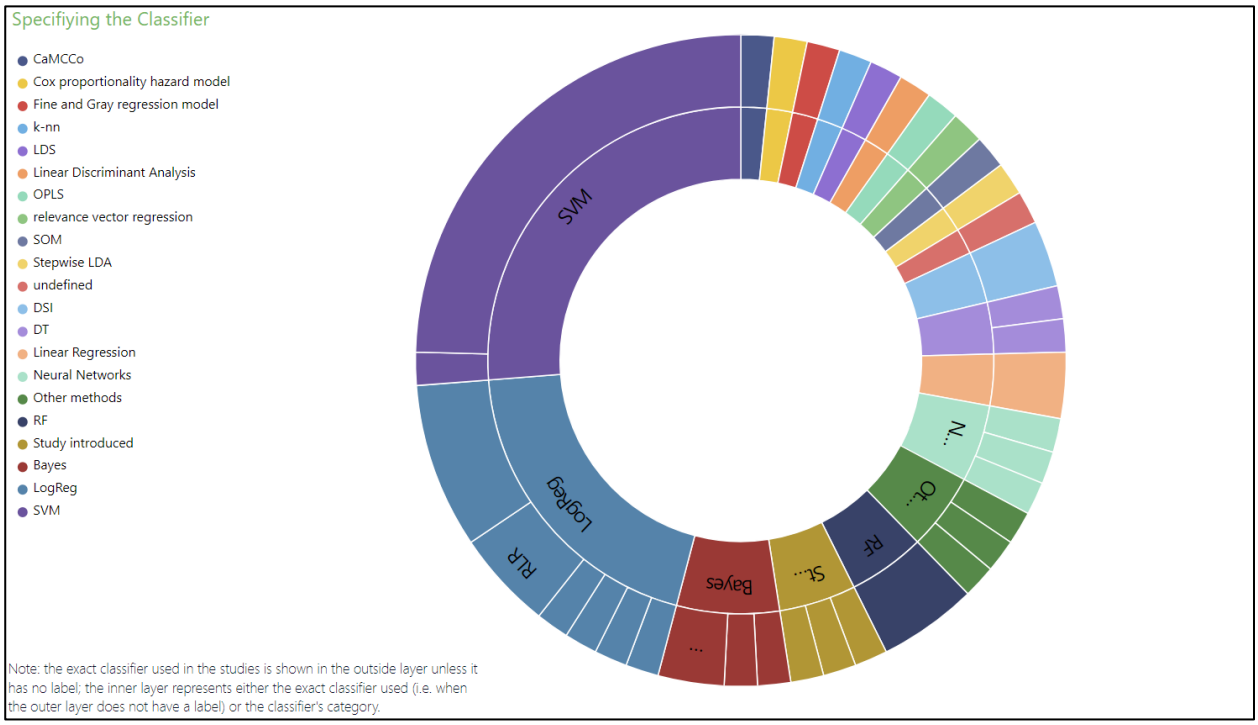


*Figure 23 - Classifier Sub Category.*

It is worth mentioning that the slots with no label mean that the same name as the parent node is applied – this is due to the existence of many models that are derived from the same principle, e.g., regularized logistic regression has its root in the logistic regression. Therefore, it was decided to leave *nulls* in the children nodes in which the model used was, in fact, the one in the parent node, as there is no further classification or specification of the models used.

*8 - Assessment Methods*

For a correct assessment of the model created, upgraded or applied, the right metric to evaluate it is mandatory.

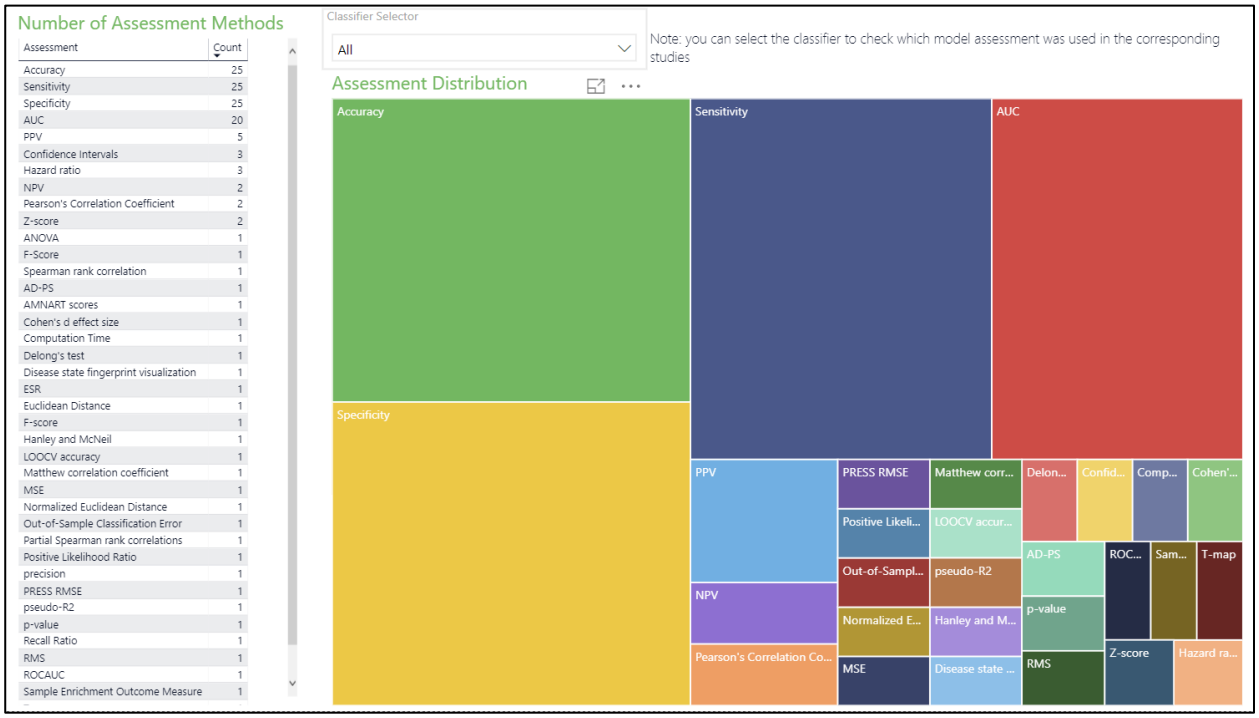


*Figure 24 - Assessment Methods.*

This dashboard aims to show the distribution of the usage of those metrics. The ones depending on the ROC curve, such as the Accuracy, Sensitivity, Specificity and Area Under the Roc curve are the most used ones. This depends greatly on the model used and the main objective of the study, as well as the data type used.

*9 - Radial Journey of Articles*

A social network analysis, like the ones briefly described in section 2.4.1, is the most suitable for representing the vast scholarly data out there, being authors names usually the main anchors of those studies, as well as the author-generated keywords. As this work does not focus on that data but on the unstructured data about the content itself instead, a simple chart was created to depict the flow between publisher, publication and article. Once got the article reference from other dashboards, it is possible to get to the publication where a specific article was edited, for instance.

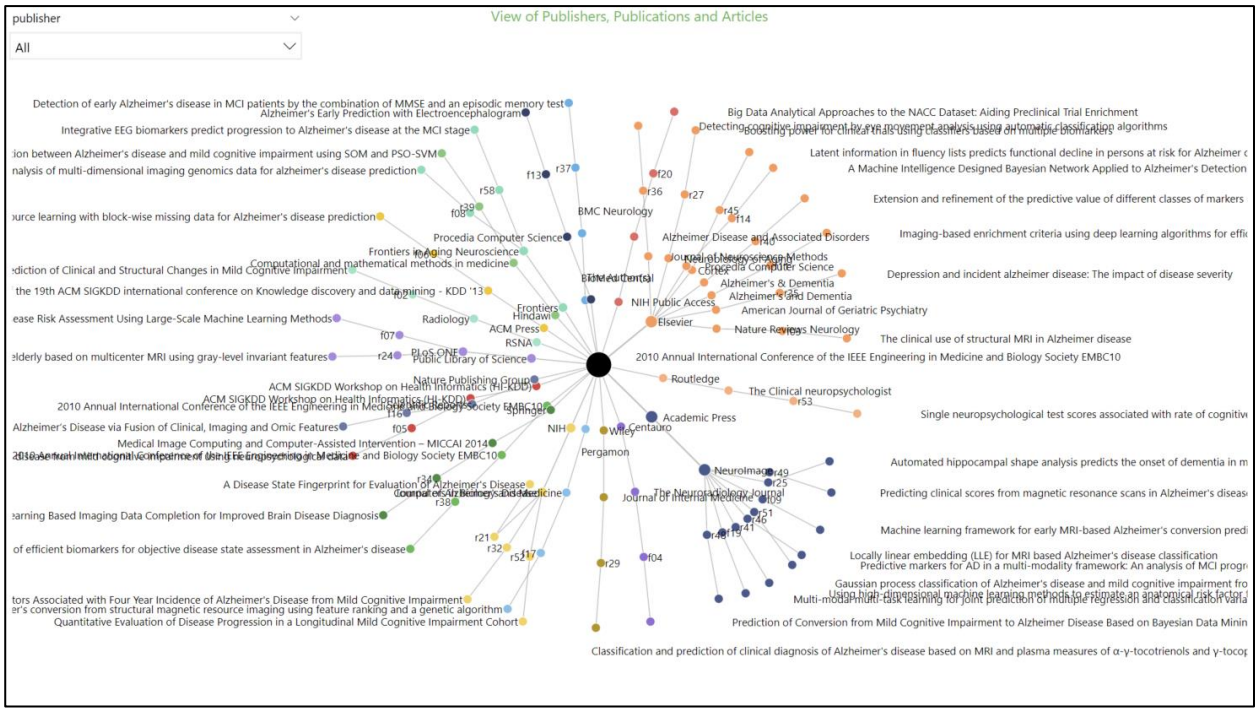


*Figure 25 - Radial Journey of Articles.*

## *10 - List of Papers*

| article_code | title | publisher | publication | year |
|---|---|---|---|---|
| f02 | Alzheimer Disease: Quantitative Structural Neuroimaging for Detection and Prediction of Clinical and Structural Changes in Mild Cognitive Impairment | RSNA | Radiology | 2009 |
| f03 | The clinical use of structural MRI in Alzheimer disease | Elsevier | Nature Reviews Neurology | 2010 |
| f04 | Prediction of Conversion from Mild Cognitive Impairment to Alzheimer Disease Based on Bayesian Data Mining with Ensemble Learning | Centauro | The Neuroradiology Journal | 2012 |
| f05 | Discriminating Alzheimer's disease from mild cognitive impairment using neuropsychological data | ACM SIGKDD Workshop on Health Informatics (HI-KDD) | ACM SIGKDD Workshop on Health Informatics (HI-KDD) | 2012 |
| f06 | Multi-source learning with block-wise missing data for Alzheimer's disease prediction | ACM Press | Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '13 | 2013 |
| f07 | Alzheimer's Disease Risk Assessment Using Large-Scale Machine Learning Methods | Public Library of Science | PLoS ONE | 2013 |
| f08 | Integrative analysis of multi-dimensional imaging genomics data for alzheimer's disease prediction | Frontiers | Frontiers in Aging Neuroscience | 2014 |
| f09 | Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects | Academic Press | NeuroImage | 2015 |
| f11 | Imaging-based enrichment criteria using deep learning algorithms for efficient clinical trials in mild cognitive impairment | Elsevier | Alzheimer's and Dementia | 2015 |
| f13 | Alzheimer's Early Prediction with Electroencephalogram | The Author(s) | Procedia Computer Science | 2016 |
| f14 | A Machine Intelligence Designed Bayesian Network Applied to Alzheimer's Detection Using Demographics and Speech Data | Elsevier | Procedia Computer Science | 2016 |
| f16 | Cascaded Multi-view Canonical Correlation (CaMCCo) for Early Diagnosis of Alzheimer's Disease via Fusion of Clinical, Imaging and Omic Features | Nature Publishing Group | Scientific Reports | 2017 |
| f17 | Classification of Alzheimer's disease and prediction of mild cognitive impairment-to-Alzheimer's conversion from structural magnetic resource imaging using feature ranking and a genetic algorithm | Pergamon | Computers in Biology and Medicine | 2017 |
| f19 | Using high-dimensional machine learning methods to estimate an anatomical risk factor for Alzheimer's disease across imaging databases | Academic Press | NeuroImage | 2018 |
| f20 | Big Data Analytical Approaches to the NACC Dataset: Aiding Preclinical Trial Enrichment | NIH Public Access | Alzheimer Disease and Associated Disorders | 2018 |
| r21 | A Disease State Fingerprint for Evaluation of Alzheimer's Disease | NIH | Journal of Alzheimer's Disease | 2011 |
| r24 | An efficient approach for differentiating Alzheimer's disease from normal elderly based on multicenter MRI using gray-level invariant features | Public Library of Science | PLoS ONE | 2014 |
| r25 | Automated hippocampal shape analysis predicts the onset of dementia in mild cognitive impairment | Academic Press | NeuroImage | 2011 |
| r27 | Boosting power for clinical trials using classifiers based on multiple biomarkers | Elsevier | Neurobiology of Aging | 2010 |
| r29 | Classification and prediction of clinical diagnosis of Alzheimer's disease based on MRI and plasma measures of α-/γ-tocotrienols and γ-tocopherol | Wiley/Blackwell (10.1111) | Journal of Internal Medicine | 2013 |

This dashboard is nothing more than a list of the articles in the database. Here one can find every article from which all the data was collected. On the right-hand side, there is a column with the year of publication of each article and it has conditional formatting so that the reader can quickly identify the most recent ones.

### 4.2.2. **Storage and Public Availability**

The visualisation platform will be available for public use at the online Power BI public library *(*Power BI App – Office 365):

https://app.powerbi.com/view?r=eyJrIjoiYWE4Njk5YmUtOGMzOC00ZDNhLTlhNWUtMTliODc5N2Y1YzFlIiwidCI6ImU0YmQ2OWZmLWU2ZjctNGMyZS1iMjQ3LTQxYjU0YmEyNDkwZSIsImMiOjh9

# 5. CONCLUSION

The era of exponential growth of digital data has only begun during this century. However, soon the topic of how to manage such amount of data and how to get information, and knowledge, out of it was raised. Soon, we learned that society would benefit if there were information systems, alongside with the right practices and quality methodologies, capable of capturing, transforming, managing and presenting data in useful ways, that would trigger action and innovation. Information systems capable of actionable information management quickly became fundamental in every sector of the world's economy; the hope of a faster and broader knowledge acquisition grew with it, and so "big data" and the other data science related subjects.

Until recently, the method followed on knowledge management has been on the acquisition and storage side. It is fair, though, since there was the need to create content before everything else, which acted as a foundation for what would come next. Now, it is time to consolidate, integrate and aggregate in order to extract the real value from that knowledge, to tackle the big problems of our society, to use the power of our collective intelligence. Organisations need that combination of actions to foster innovation, collaboration and value creation. It is also true for the scientific research and for the scholarly data in general, from which a tremendous amount of valuable content is generated in the form or scientific publication, but not stored, shared or integrated according to the latest knowledge management practices to be reused and to enable an accelerated knowledge discovery. Surprisingly, despite the proliferation of publication websites, conferences, online journals, and so forth, scientific publications are still spread throughout the world wide web and stored, mainly, in commercial publication storage websites, making it extremely hard to keep track of the recent development in any scientific area. Even today, after so much evolution on the information systems, there is no single, unified, integrated system which has all the information and knowledge gathered from all the studies performed so far. Thus, it is becoming imperative to have a more integrative information and knowledge management system with the goal of providing quick and reliable identification of the scientific articles and developments in such sensitive and vital areas such as Alzheimer's Disease.

With it in mind, this dissertation narrows tackles that identified a gap in the knowledge management process, towards and advance knowledge sharing and collaboration between organisation, institutions, rookies and experts from the different field that is involved with the AD's research. A framework for systematic information management to keep track of the evolutions as far as the application of predictive analytics on Alzheimer's Disease discovery was developed. It was performed a systematic scientific publication review and analysis to portrait the current state of the art on the topic. That analysis preceded a broad search for scientific articles on the online reference scientific publications databases. Then, the relevant content, namely the research objective, the data types and source, the data pre-processing, data modelling and model assessment techniques, methods and approaches were manually extracted from those articles – the ones under the scope of this work - and categorised so that it could be stored in a relational start-schema shaped database, using Microsoft Excel and Power BI Desktop. With the collected data of fifty-eight articles, a data visualisation framework was developed in Power BI Desktop to unveil what has been the role of predictive analytics in this field of scientific research. For this framework, twelve interactive dashboards were developed to make it as clear as possible the current state of the art proposed by this dissertation. From the data and visual analysis performed, it became clear that the most used dataset source is ADNI and that the neuroimaging data is the most used data type in those studies,

in particular, Magnetic Resonance Imaging. This database is commonly used by the researchers that are studying the prediction of MCI to AD conversion or the trying to improve the diagnosis of AD or MCI as the primary objective for their study. When it comes to the mostly used classifiers, SVM classifier and the Logistic Regression classifiers category are the most used ones; those are usually associated with the usage of neuropsychological and neuroimaging data. Concerning the data pre-processing stage of the Knowledge Data Discovery Process, a broad range of techniques have been used, whereas accuracy, specificity and sensitivity are the most common metrics to evaluate the performance of the model created of used in the study.

With this work, it became even more evident that it is challenging not only to keep track of all the studies that have been done in the world about a specific topic but also how difficult it is to get relevant data from articles' content due to the heterogeneous nature of the published articles. It creates, thus, a considerable barrier to go from unstructured to structured data, even more, when the knowledge on the specific scientific research field is not high. This does not only happen in the Alzheimer's Disease Research but it is trans sectorial to many other research topics.

The claim, introduced by this dissertation that it is rather difficult to track every single published article, was proven to be accurate during the execution of it. It was made clear as well that it takes only a quick subsequent search for articles on the subject addressed by this dissertation, or by analysing the references of articles already found, that is rather easy to miss relevant on-topic documents that could have been added to the database developed in this work. It is, therefore, and undeniable evidence that the information management practices in the scientific community need to be improved, perhaps reformed, to become more efficient, more transparent and genuinely collaborative.

Furthermore, it would benefit many stakeholders, from the researchers in the scientific community to educational institutions to decision makers on a governmental level, who could potentially better manage the research funding allocation and decide on the practical implementation of some of the techniques already developed.

# 6. LIMITATIONS AND FUTURE WORK

This work has, inevitably, some limitations.

Firstly, there are access limitations found in some online scientific resources, in which a payment is required to read the article, have limited the number of articles gathered. Secondly, the visualisation tool only presents the number based on the articles that were possible to collect. Thirdly, and continuing on the data visualisation framework topic, it could have been useful to develop it using entirety using JavaScript with the d3.js library as it would increase not only the flexibility of the tool regarding UX and UI (supposedly) but it would give more freedom as far as design choices. With that, a new and unique website could be developed using the framework developed in this work and presenting the data in ways it has not been seen before.

Therefore, and tacking the ideas for possible future work, some interesting further developments can be applied. To start with, the usage of a text mining tool to retrieve all the necessary topics automatically from the articles' content would speed up the process and provide a genuinely E2E information management solution. It is crucial, for the purpose at hands, to have machine learning-based tools that help on the transformation process of structuring data from text, as easy and fast as possible.

Moreover, even automatically update the database and, consequently, automating the update of the data visualisation tool. The system, then, is expected to have regular updates, collecting and presenting an always updated version of the state-of-science about a holistic and integrative big data approach on the prediction of Alzheimer's disease. Furthermore, Hypothesis Finder could be a solution to get content data

If the time dimension is implemented in the relation database, a whole new level of data exploration could be unleashed; an OLAP cube development should also be considered for better trend tracking, e.g. on the increase or decrease trend of the usage of certain classifiers.

One other idea would be to start integrating the social network analysis with the article's content database is the creating of fields for the articles keywords as well as the authors, as the majority of the scholarly social network analysis focus on these two domains. It is probably an important aspect as the reader may want to increase further his/her knowledge in the area. Then, he/she will try to find related work through the references and bibliography. Therefore, it is essential to have a proper scholarly network system established to speed up this information retrieval.

Lastly, it would be interesting to extend this analytical framework for other important scientific research fields, not only in the neurodegenerative diseases field, such as Parkinson's disease, but also to other medicine related fields, where data is becoming more and more available, such as rare forms of cancer.

## BIBLIOGRAPHY AND REFERENCES


Alzheimer's Association. (2015). 2015 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, *11*(3), 332–384. https://doi.org/10.1016/j.jalz.2015.02.003

Alzheimer's Association. (2016). Alzheimer's & Dementia Testing Advances | Research Center | Alzheimer's Association. Retrieved from http://www.alz.org/research/science/earlier_alzheimers_diagnosis.asp

Association, A. (2017). 2017 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, *13*(4), 325–373. https://doi.org/10.1016/j.jalz.2017.02.001

Ballard, C. C. C., Gauthier, S., Corbett, A., Brayne, C., Aarsland, D., Jones, E., … Neurology. (2011). Alzheimer's disease. *Lancet (London, England)*, *377*(9770), 1019–31. https://doi.org/10.1016/s0140-6736(10)61349-9

Barrett, M. A., Humblet, O., Hiatt, R. A., & Adler, N. E. (2013). Big Data and Disease Prevention: *From Quantified Self to Quantified Communities*. *Big Data*, *1*(3), 168–175. https://doi.org/10.1089/big.2013.0027

Beheshti, I., Demirel, H., & Matsuda, H. (2017). Classification of Alzheimer's disease and prediction of mild cognitive impairment-to-Alzheimer's conversion from structural magnetic resource imaging using feature ranking and a genetic algorithm. *Computers in Biology and Medicine*, *83*, 109–119. https://doi.org/10.1016/j.compbiomed.2017.02.011

Belle, A. A. ., Thiagarajan, R. R. ., Soroushmehr, S. M. R. . M. R., Navidi, F. F. . F., Beard, D. a. D. A. . D. a. D. A. ., & Najarian, K. K. . (2015). Big Data Analytics in Healthcare. *BioMed Research International*, *2015*(8), 1–16. https://doi.org/10.1155/2015/370194

Bhavnani, S. P., Narula, J., & Sengupta, P. P. (2016). Mobile technology and the digitization of healthcare. *European Heart Journal*, *37*(18), 1428–1438. https://doi.org/10.1093/eurheartj/ehv770

Bill Gates. (2017). Why I'm digging deep into Alzheimer's | Bill Gates. Retrieved July 2, 2018, from https://www.gatesnotes.com/Health/Digging-Deep-Into-Alzheimers

Bochicchio, M., Cuzzocrea, A., & Vaira, L. (2016). A Big Data Analytics Framework for Supporting Multidimensional Mining over Big Healthcare Data. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 508–513). IEEE. https://doi.org/10.1109/ICMLA.2016.0090

Bubu, O., Bakke, J., Hogan, M., Umasabor-Bubu, O., Mukhtar, F., Ram, S., & Osorio, R. (2017). 1153 Disturbed Sleep Is Associated With Changes In Alzheimer's Disease (Ad) Biomarkers Predictive Of Persons That Ultimately Develop Ad: Findings From Subgroup Meta-Analysis On Sleep And Alzheimer's Disease. *Sleep*, *40*(suppl_1), A430–A430. https://doi.org/10.1093/sleepj/zsx050.1152

Cano, I., Tenyi, A., Vela, E., Miralles, F., & Roca, J. (2017). Perspectives on Big Data applications of health information. *Current Opinion in Systems Biology*, *3*, 36–42. https://doi.org/10.1016/j.coisb.2017.04.012

Caragea, C., Wu, J., Williams, K., Gollapalli, S. Das, Khabsa, M., Teregowda, P., & Giles, C. L. (2014). Automatic identification of research articles from crawled documents. In *WSDM 2014 Workshop on Web-scale Classification: Classifying Big Data from the Web*.

Carreiro, A. V., Mendonça, A., de Carvalho, M., & Madeira, S. C. (2015). Integrative biomarker discovery

in neurodegenerative diseases. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, *7*(6), 357–379. https://doi.org/10.1002/wsbm.1310

Casanova, R., Barnard, R. T., Gaussoin, S. A., Saldana, S., Hayden, K. M., Manson, J. E., … Chen, J.-C. (2018). Using high-dimensional machine learning methods to estimate an anatomical risk factor for Alzheimer's disease across imaging databases. *NeuroImage*, *183*, 401–411. https://doi.org/10.1016/j.neuroimage.2018.08.040

Chen, R., Young, K., Chao, L. L., Miller, B., Yaffe, K., Weiner, M. W., & Herskovits, E. H. (2012). Prediction of conversion from mild cognitive impairment to Alzheimer disease based on Bayesian data mining with ensemble learning. *The Neuroradiology Journal*, *25*(1), 5–16. https://doi.org/10.1177/197140091202500101

Chen, Y., Argentinis, E., & Weber, G. (2016). IBM Watson: How Cognitive Computing Can Be Applied to Big Data Challenges in Life Sciences Research. *Clinical Therapeutics*. https://doi.org/10.1016/j.clinthera.2015.12.001

citespace101. (2016). citespace101. Retrieved from https://sites.google.com/site/citespace101/home?authuser=0

Dinov, I. D., Heavner, B., Tang, M., Glusman, G., Chard, K., Darcy, M., … Toga, A. W. (2016). Predictive big data analytics: A study of Parkinson's disease using large, complex, heterogeneous, incongruent, multi-source and incomplete observations. *PLoS ONE*, *11*(8), e0157077. https://doi.org/10.1371/journal.pone.0157077

Evergreen, S. (2016). *Effective data visualization : the right chart for the right data*. Retrieved from https://books.google.pt/books?hl=en&lr=&id=PICzDAAAQBAJ&oi=fnd&pg=PP1&dq=data+visualization+in+alzheimer&ots=_emUi9Fpgs&sig=XJ3fDwP136zsHouMW4dlHUr5Q9Y&redir_esc=y#v=onepage&q=communicate&f=false

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37–37. https://doi.org/10.1609/AIMAG.V17I3.1230

Frisoni, G. B., Fox, N. C., Jack, C. R., Scheltens, P., & Thompson, P. M. (2010). The clinical use of structural MRI in Alzheimer disease. *Nature Reviews Neurology*, *6*(2), 67–77. https://doi.org/10.1038/nrneurol.2009.215

Fung, T. L. (2015). Visual Characterization of Personal Bibliographic Data Using a Botanical Tree Design. *In Proceedings of IEEE VIS 2015 Workshop on Personal Visualization*, 2–5.

Geerts, H., Dacks, P. A., Devanarayan, V., Haas, M., Khachaturian, Z. S., Gordon, M. F., … Stephenson, D. (2016). Big data to smart data in Alzheimer's disease: The brain health modeling initiative to foster actionable knowledge. *Alzheimer's and Dementia*. Elsevier Inc. https://doi.org/10.1016/j.jalz.2016.04.008

Haas, M., Stephenson, D., Romero, K., Gordon, M. F., Zach, N., & Geerts, H. (2016). Big data to smart data in Alzheimer's disease: Real-world examples of advanced modeling and simulation. *Alzheimer's & Dementia*, *12*(9), 1022–1030. https://doi.org/10.1016/j.jalz.2016.05.005

Habes, M., Janowitz, D., Erus, G., Toledo, J. B., Resnick, S. M., Doshi, J., … Davatzikos, C. (2016). Advanced brain aging: Relationship with epidemiologic and genetic risk factors, and overlap with Alzheimer disease atrophy patterns. *Translational Psychiatry*, *6*(4), e775. https://doi.org/10.1038/tp.2016.39

Hagerty, J., Sallam, R. L., & Richardson, J. (2011). Magic Quadrant for Business Intelligence Platforms.

*Gartner for Business …*, (February), 1–52. https://doi.org/G00239854

Healey, C. G. (2007). Perception in visualization. *Retrieved February*, *10*(June), 2008. Retrieved from https://www.csc2.ncsu.edu/faculty/healey/PP/index.html

Healy, K. (2018). *Data Visualization: A practical introduction* (2018th-04-25th ed.). Forthcoming, Princeton University Press. Retrieved from http://socviz.co/index.html#preface

Histcite. (2018). Histcitee. Retrieved from https://en.wikipedia.org/wiki/Histcite

Hofmann-Apitius, M. (2015). Is dementia research ready for big data approaches? *BMC Medicine*, *13*(1), 145. https://doi.org/10.1186/s12916-015-0367-7

Hoyt, R. E., Snider, D., Thompson, C., & Mantravadi, S. (2016). IBM Watson Analytics: Automating Visualization, Descriptive, and Predictive Statistics. *JMIR Public Health and Surveillance*, *2*(2), e157. https://doi.org/10.2196/publichealth.5810

Huang, T., Lan, L., Fang, X., An, P., Min, J., & Wang, F. (2015). Promises and Challenges of Big Data Computing in Health Sciences. *Big Data Research*. https://doi.org/10.1016/j.bdr.2015.02.002

Ipp, C. I., Azevedo, A., & Santos, M. F. (2004). KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW. *Appliance*, *61*(1), 35.

Isenberg, P., Isenberg, T., Sedlmair, M., Chen, J., & Möller, T. (2017). Visualization as Seen through its Research Paper Keywords. *IEEE Transactions on Visualization and Computer Graphics*, *23*(1), 771–780. https://doi.org/10.1109/TVCG.2016.2598827

Ithapu, V. K., Singh, V., Okonkwo, O. C., Chappell, R. J., Dowling, N. M., & Johnson, S. C. (2015). Imaging-based enrichment criteria using deep learning algorithms for efficient clinical trials in mild cognitive impairment. *Alzheimer's and Dementia*, *11*(12), 1489–1499. https://doi.org/10.1016/j.jalz.2015.01.010

Janson, J., Laedtke, T., Parisi, J. E., O'Brien, P., Petersen, R. C., & Butler, P. C. (2004). Increased Risk of Type 2 Diabetes in Alzheimer Disease. *Diabetes*, *53*(2), 474–481. https://doi.org/10.2337/diabetes.53.2.474

Jiang, X., & Zhang, J. (2016). A text visualization method for cross-domain research topic mining. *Journal of Visualization*, *19*(3), 561–576. https://doi.org/10.1007/s12650-015-0323-9

Kantar. (2018). *Information is Beautiful Awards*. Retrieved from https://www.informationisbeautifulawards.com/

Kaur, H., & Wasan, S. K. (2006). Empirical Study on Applications of Data Mining Techniques in Healthcare. *(B3) Journal of Computer Sciences*, *2*(2), 194–200. https://doi.org/10.3844/jcssp.2006.194.200

Koh, H. C., & Tan, G. (2005). Data mining applications in healthcare. *J Healthc Inf Manag*, *19*(2), 64–72. https://doi.org/10.4314/ijonas.v5i1.49926

Krippendorff, K. (1980). *Content analysis : an introduction to its methodology* (1st ed.). Retrieved from https://uk.sagepub.com/en-gb/eur/content-analysis/book258450

Land, W. H., & Schaffer, J. D. (2016). A Machine Intelligence Designed Bayesian Network Applied to Alzheimer's Detection Using Demographics and Speech Data. *Procedia Computer Science*. https://doi.org/10.1016/j.procs.2016.09.308

Lemos, L. J. M., Silva, D., Guerreiro, M., Santana, I., Mendonça, A., Tomás, P., & Madeira, S. C. (2012). Discriminating Alzheimer's disease from mild cognitive impairment using neuropsychological data. *ACM SIGKDD Workshop on Health Informatics (HI-KDD)*.

Li, R., Zhang, W., Suk, H.-I., Wang, L., Li, J., Shen, D., & Ji, S. (2014). Deep Learning Based Imaging Data Completion for Improved Brain Disease Diagnosis (pp. 305–312). Springer, Cham. https://doi.org/10.1007/978-3-319-10443-0_39

Lin, M., Gong, P., Yang, T., Ye, J., Albin, R. L., & Dodge, H. H. (2018). Big Data Analytical Approaches to the NACC Dataset: Aiding Preclinical Trial Enrichment. *Alzheimer Disease and Associated Disorders*, *32*(1), 18–27. https://doi.org/10.1097/WAD.0000000000000228

Liu, J., Tang, T., Wang, W., Xu, B., Kong, X., & Xia, F. (2018). A Survey of Scholarly Data Visualization. *IEEE Access*, *6*, 19205–19221. https://doi.org/10.1109/ACCESS.2018.2815030

Malik, M. M., Abdallah, S., & Ala'raj, M. (2018). Data mining and predictive analytics applications for the delivery of healthcare services: a systematic literature review. *Annals of Operations Research*, *270*(1–2), 287–312. https://doi.org/10.1007/s10479-016-2393-z

Maudsley, S., Devanarayan, V., Martin, B., & Geerts, H. (2018). Intelligent and effective informatic deconvolution of "Big Data" and its future impact on the quantitative nature of neurodegenerative disease therapy. *Alzheimer's and Dementia*, *14*(7), 961–975. https://doi.org/10.1016/j.jalz.2018.01.014

McCandless, D. (2012). *Information Is Beautiful*. *Information is beautiful*. https://doi.org/10.1016/j.arth.2014.09.018

McEvoy, L. K., Fennema-Notestine, C., Roddey, J. C., Hagler, D. J., Holland, D., Karow, D. S., … Alzheimer's Disease Neuroimaging Initiative, A. M. (2009). Alzheimer disease: quantitative structural neuroimaging for detection and prediction of clinical and structural changes in mild cognitive impairment. *Radiology*, *251*(1), 195–205. https://doi.org/10.1148/radiol.2511080924

Moradi, E., Pepe, A., Gaser, C., Huttunen, H., & Tohka, J. (2015). Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *NeuroImage*, *104*, 398–412. https://doi.org/10.1016/j.neuroimage.2014.10.002

Munzner, T., & Maguire, E. (Graphic artist). (2014). *Visualization analysis and design* (1st ed.). A K Peters/CRC Press. Retrieved from https://books.google.pt/books/about/Visualization_Analysis_and_Design.html?id=4eqsBAAAQBAJ&source=kp_cover&redir_esc=y

Nagpal, S., Arora, S., Dey, S., & Shreya, S. (2017). Feature Selection using Gravitational Search Algorithm for Biomedical Data. *Procedia Computer Science*, *115*, 258–265. https://doi.org/10.1016/j.procs.2017.09.133

Ojha, M., & Mathur, K. (2016). Proposed application of big data analytics in healthcare at Maharaja Yeshwantrao Hospital. In *2016 3rd MEC International Conference on Big Data and Smart City, ICBDSC 2016* (pp. 40–46). https://doi.org/10.1109/ICBDSC.2016.7460340

Ortiz, A., Munilla, J., Górriz, J. M., & Ramírez, J. (2016). Ensembles of Deep Learning Architectures for the Early Diagnosis of the Alzheimer's Disease. *International Journal of Neural Systems*, *26*(07), 1650025. https://doi.org/10.1142/S0129065716500258

Pereira, T., Lemos, L., Cardoso, S., Silva, D., Rodrigues, A., Santana, I., … Madeira, S. C. (2017). Predicting progression of mild cognitive impairment to dementia using neuropsychological data: A supervised

learning approach using time windows. *BMC Medical Informatics and Decision Making*, *17*(1), 1–15. https://doi.org/10.1186/s12911-017-0497-2

Perera, G., Pedersen, L., Ansel, D., Alexander, M., Arrighi, H. M., Avillach, P., … Stewart, R. (2018). Dementia prevalence and incidence in a federation of European Electronic Health Record databases: The European Medical Informatics Framework resource. *Alzheimer's and Dementia*, *14*(2), 130–139. https://doi.org/10.1016/j.jalz.2017.06.2270

Pham, T., Tran, T., Phung, D., & Venkatesh, S. (2016). DeepCare: A Deep Dynamic Memory Model for Predictive Medicine. Retrieved from http://arxiv.org/abs/1602.00357

Pham, T., Tran, T., Phung, D., & Venkatesh, S. (2017). Predicting healthcare trajectories from medical records: A deep learning approach. *Journal of Biomedical Informatics*, *69*, 218–229. https://doi.org/10.1016/j.jbi.2017.04.001

Prather, J. C., Lobach, D. F., Goodwin, L. K., Hales, J. W., Hage, M. L., & Hammond, W. E. (1997). Medical data mining: knowledge discovery in a clinical data warehouse. *Proceedings : A Conference of the American Medical Informatics Association. AMIA Fall Symposium*, 101–5. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9357597

Prince, M., Comas-Herrera, A., Knapp, M., Guerchet, M., & Karagiannidou, M. (2016). World Alzheimer Report 2016 Improving healthcare for people living with dementia. Coverage, Quality and costs now and in the future. *Alzheimer's Disease International (ADI)*, 1–140. https://doi.org/10.13140/RG.2.2.22580.04483

Qureshi, S., Briggs, R. O., & Hlupic, V. (2006). Value Creation from Intellectual Capital: Convergence of Knowledge Management and Collaboration in the Intellectual Bandwidth Model. *Group Decision and Negotiation*, *15*(3), 197–220. https://doi.org/10.1007/s10726-006-9018-x

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, *2*(1), 3. https://doi.org/10.1186/2047-2501-2-3

Ravi, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., & Yang, G.-Z. Z. (2016). Deep Learning for Health Informatics. *IEEE Journal of Biomedical and Health Informatics*, *21*(1), 4–21. https://doi.org/10.1109/JBHI.2016.2636665

Rodrigues, P. M., Teixeira, J. P., Garrett, C., Alves, D., & Freitas, D. (2016). Alzheimer's Early Prediction with Electroencephalogram. *Procedia Computer Science*, *100*, 865–871. https://doi.org/10.1016/j.procs.2016.09.236

Sakr, S., & Elgammal, A. (2016). Towards a Comprehensive Data Analytics Framework for Smart Healthcare Services. *Big Data Research*, *4*, 44–58. https://doi.org/10.1016/j.bdr.2016.05.002

Saravana Kumar, N. M., Eswari, T., Sampath, P., & Lavanya, S. (2015). Predictive methodology for diabetic data analysis in big data. In *Procedia Computer Science* (Vol. 50, pp. 203–208). https://doi.org/10.1016/j.procs.2015.04.069

sci2. (2018). sci2. Retrieved from https://sci2.cns.iu.edu/user/index.php

Serrano-Pozo, A., Frosch, M. P., Masliah, E., & Hyman, B. T. (2011). Neuropathological alterations in Alzheimer disease. *Cold Spring Harbor Perspectives in Medicine*. https://doi.org/10.1101/cshperspect.a006189

Singanamalli, A., Wang, H., Madabhushi, A., Weiner, M., Aisen, P., Petersen, R., … Fargher, K. (2017). Cascaded Multi-view Canonical Correlation (CaMCCo) for Early Diagnosis of Alzheimer's Disease via

Fusion of Clinical, Imaging and Omic Features. *Scientific Reports*, *7*(1), 8137. https://doi.org/10.1038/s41598-017-03925-0

Siuly, S., & Zhang, Y. (2016). Medical Big Data: Neurological Diseases Diagnosis Through Medical Data Analysis. *Data Science and Engineering*, *1*(2), 54–64. https://doi.org/10.1007/s41019-016-0011-3

Sukumar, S. R., Natarajan, R., & Ferrell, R. K. (2015). Quality of Big Data in health care. *International Journal of Health Care Quality Assurance*, *28*(6), 621–634. https://doi.org/10.1108/IJHCQA-07-2014-0080

Sun, J., & Reddy, C. C. K. (2013). Big data analytics for healthcare. *SIAM International Conference on Knowledge Discovery and Data …*, 1525. https://doi.org/10.1145/2487575.2506178

Tejeswinee, K., Shomona, G. J., & Athilakshmi, R. (2017). Feature Selection Techniques for Prediction of Neuro-Degenerative Disorders: A Case-Study with Alzheimer's and Parkinson's Disease. *Procedia Computer Science*. https://doi.org/10.1016/j.procs.2017.09.125

Tennant, J. (2018). Elsevier are corrupting open science in Europe. Retrieved from https://www.theguardian.com/science/political-science/2018/jun/29/elsevier-are-corrupting-open-science-in-europe

Tomar, D., & Agarwal, S. (2013). A survey on Data Mining approaches for Healthcare. *International Journal of Bio-Science and Bio-Technology*, *5*(5), 241–266. https://doi.org/10.14257/ijbsbt.2013.5.5.25

Tufte, E. R. (2001). *The visual display of quantitative information*. Graphics Press.

Vosviewer. (2018). Vosviewer. Retrieved from http://www.vosviewer.com/products

Wang, Y., & Hajli, N. (2017). Exploring the path to big data analytics success in healthcare. *Journal of Business Research*, *70*, 287–299. https://doi.org/10.1016/j.jbusres.2016.08.002

Wang, Y., Kung, L. A., & Byrd, T. A. (2015). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change*. https://doi.org/10.1016/j.techfore.2015.12.019

Wang, Y., Kung, L., Ting, C., & Byrd, T. A. (2015). Beyond a Technical Perspective: Understanding Big Data Capabilities in Health Care. In *2015 48th Hawaii International Conference on System Sciences* (Vol. 2015–March, pp. 3044–3053). IEEE. https://doi.org/10.1109/HICSS.2015.368

Wang, Y., Kung, L., Wang, W. Y. C., & Cegielski, C. G. (2017). An integrated big data analytics-enabled transformation model: Application to health care. *Information & Management*, *55*(1), 64–79. https://doi.org/10.1016/j.im.2017.04.001

World Health Organization. (2012). Dementia: a public health priority. *WHO Press*. Retrieved from http://apps.who.int/iris/bitstream/10665/75263/1/9789241564458_eng.pdf?ua=1

Yau, N. (2011). *Visualize this : the FlowingData guide to design, visualization, and statistics*. Wiley Pub. Retrieved from https://www.wiley.com/en-us/Visualize+This%3A+The+FlowingData+Guide+to+Design%2C+Visualization%2C+and+Statistics-p-9780470944882

Zhang, R., Simon, G., & Yu, F. (2017). Advancing Alzheimer's research: A review of big data promises. *International Journal of Medical Informatics*, *106*, 48–56. https://doi.org/10.1016/j.ijmedinf.2017.07.002

Zhang, Z., Huang, H., & Shen, D. (2014). Integrative analysis of multi-dimensional imaging genomics data for alzheimer's disease prediction. *Frontiers in Aging Neuroscience*, *6*(SEP), 260. https://doi.org/10.3389/fnagi.2014.00260

# ANNEX

## Extra Data Analysis

*Table 11 – Number of articles the according to their main study objective.*

| Main Objective of the study | Count |
|---|---|
| describing clinical care for persons with AD | **3** |
| Diagnosing AD or MCI | **17** |
| mining the literature for knowledge discovery | **5** |
| predicting AD progression | **3** |
| predicting AD progression. | **1** |
| predicting MCI to AD conversion | **16** |
| stratifying risks for AD | **4** |
| understanding the relationship between cognition and AD | **3** |
| **Grand Total** | **52** |

*Table 12 – Number of Items in the Analytical Process per Article.*

| Article code | #datasets | #classifiers | #Assessments | #data_prep | #dataType | #MRO |
|---|---|---|---|---|---|---|
| **f02** | 1 | 1 | 5 | 2 | 2 | 1 |
| **f03** | | 1 | | 1 | 1 | 1 |
| **f04** | 1 | 1 | 3 | 3 | 2 | 1 |
| **f05** | 1 | 5 | 5 | 5 | 1 | 1 |
| **f06** | 2 | 1 | 5 | 8 | 3 | 1 |
| **f07** | 1 | 1 | 3 | 1 | 5 | 1 |
| **f08** | 2 | 1 | 3 | 4 | 4 | 1 |
| **f09** | 1 | 3 | 4 | 1 | 3 | 1 |
| **f11** | 1 | 1 | 1 | 1 | 9 | 1 |
| **f13** | 1 | 1 | 5 | 1 | 2 | 1 |
| **f14** | 1 | 3 | 4 | 1 | 3 | 1 |
| **f16** | 1 | 1 | 5 | 1 | 7 | 1 |
| **f17** | 1 | 1 | 4 | 3 | 1 | 1 |
| **f19** | 1 | 1 | 1 | 1 | 1 | 1 |
| **f20** | 1 | 3 | 4 | 10 | 1 | 1 |
| **r21** | 1 | 4 | 1 | 1 | 7 | 1 |
| **r24** | 1 | 1 | 5 | 1 | 1 | 1 |
| **r25** | 2 | 1 | 5 | 1 | 4 | 1 |
| **r27** | 1 | 1 | 2 | 1 | 1 | 1 |
| **r29** | 1 | 1 | 5 | 1 | 2 | 1 |
| **r32** | 1 | 2 | 3 | 2 | 3 | 1 |
| **r34** | 1 | 1 | 1 | 2 | 1 | 1 |
| **r35** | 1 | 2 | 2 | 1 | 3 | 1 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **r36** | 1 | 2 | 4 | 1 | 1 | 1 |
| **r37** | 1 | 1 | 5 | 1 | 5 | 1 |
| **r38** | 2 | 2 | 1 | 3 | 14 | 1 |
| **r39** | 1 | 3 | 3 | 2 | 3 | 1 |
| **r40** | 1 | 1 | 5 | 3 | 8 | 1 |
| **r41** | 1 | 1 | 4 | 3 | 2 | 1 |
| **r45** | 1 | 1 | 2 | 1 | 4 | 2 |
| **r46** | 1 | 3 | 4 | 1 | 1 | 1 |
| **r48** | 1 | 1 | 2 | 1 | 4 | 1 |
| **r49** | 2 | 1 | 2 | 3 | 4 | 1 |
| **r51** | 1 | 3 | 4 | 1 | 5 | 1 |
| **r52** | 1 | 2 | 1 | 1 | 4 | 1 |
| **r53** | 1 | 1 | 4 | 1 | 4 | 1 |
| **r58** | 1 | 1 | 5 | 2 | 3 | 1 |

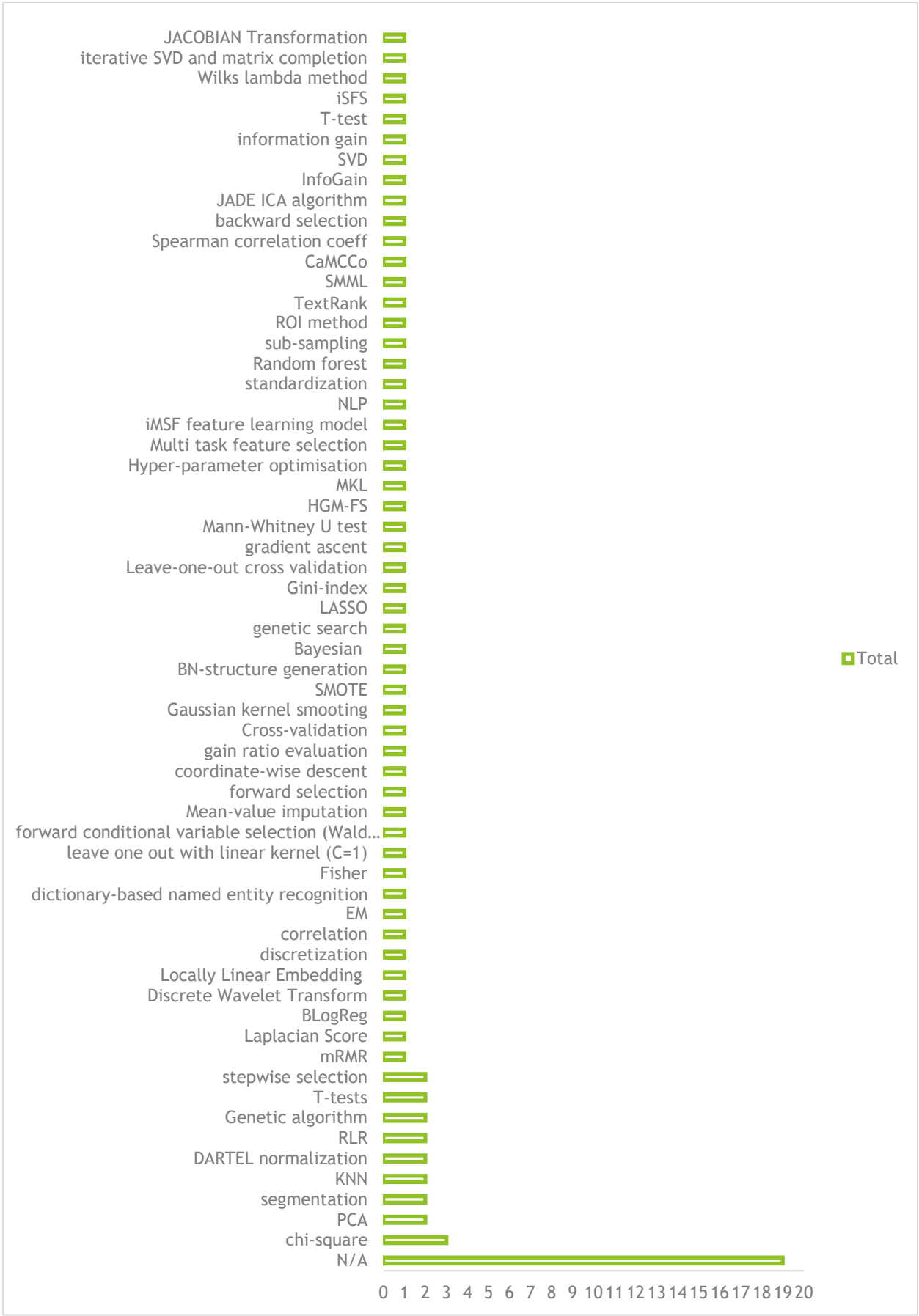


*Figure 26 – List of Data Pre-Processing methods used in the studies found.*

## Power BI Implementation

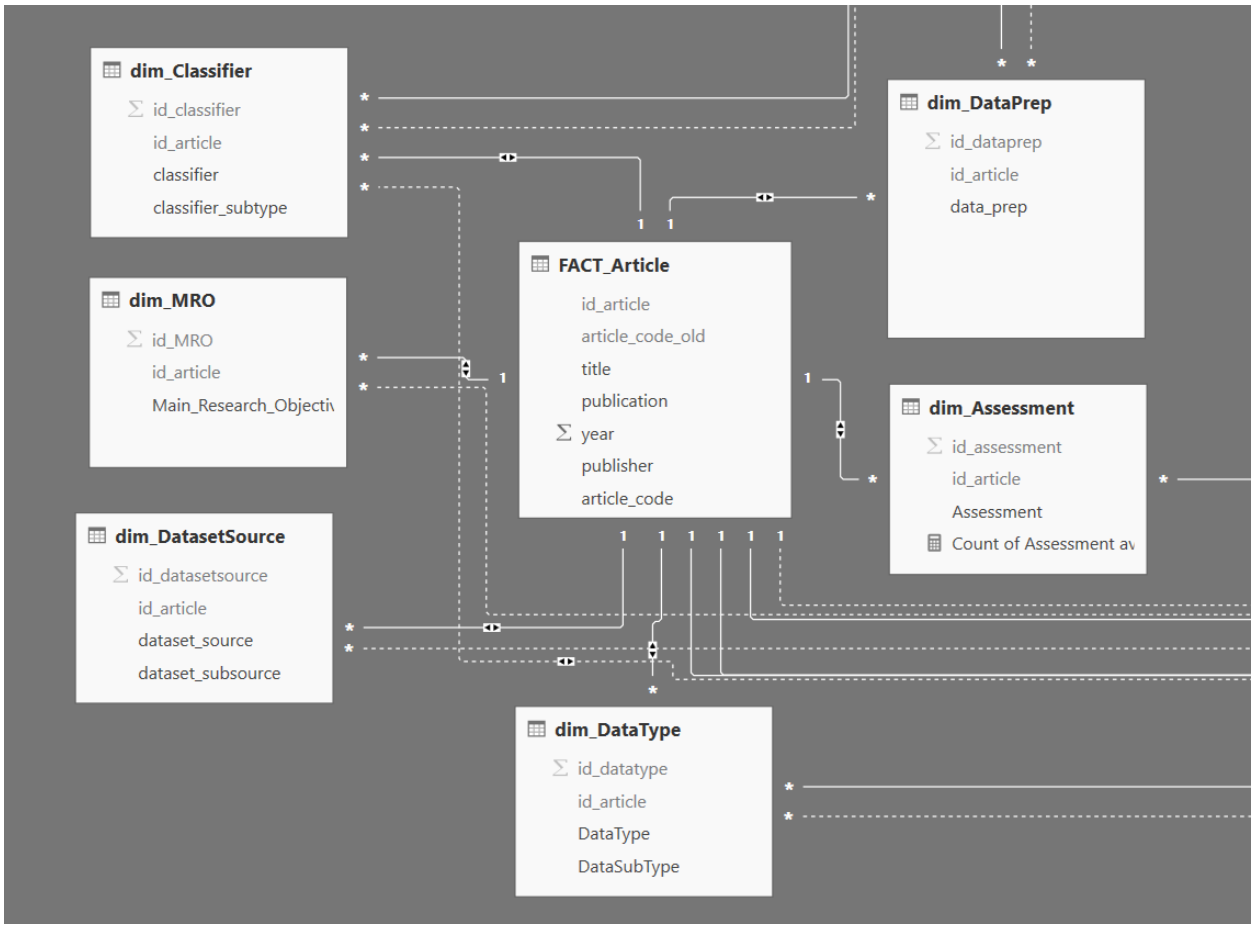


*Figure 27 – Implementation of the relational database in Power BI.*

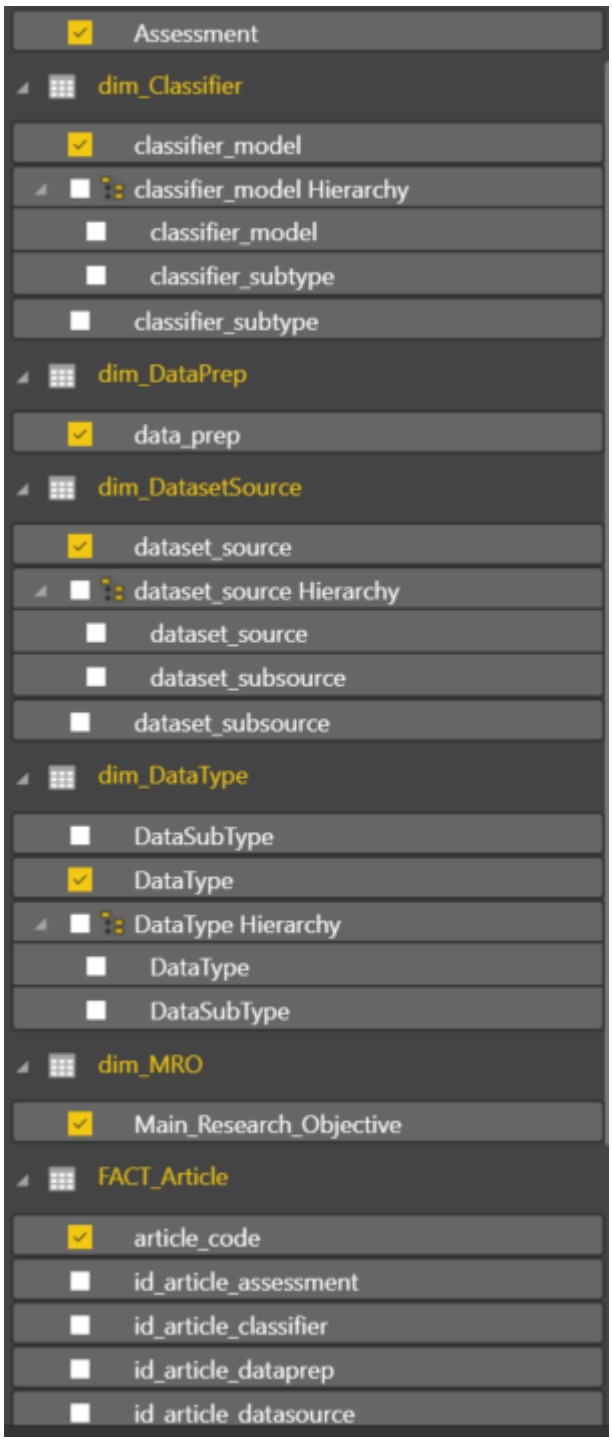


*Figure 28 –Tables in the Relational Database and Hierarchies in Power BI.*

## Complete List of Articles

*Table 13 - Complete List of Articles Included in the Database.*

| Article Code | Title | Publication | Year | publisher |
|---|---|---|---|---|
| **f2** | Alzheimer Disease: Quantitative Structural Neuroimaging for Detection and Prediction of Clinical and Structural Changes in Mild Cognitive Impairment | Radiology | 2009 | RSNA |
| **f3** | The clinical use of structural MRI in Alzheimer disease | Nature Reviews Neurology | 2010 | Elsevier |
| **f4** | Prediction of Conversion from Mild Cognitive Impairment to Alzheimer Disease Based on Bayesian Data Mining with Ensemble Learning | The Neuroradiology Journal | 2012 | Centauro |
| **f5** | Discriminating Alzheimer's disease from mild cognitive impairment using neuropsychological data | ACM SIGKDD Workshop on Health Informatics (HI-KDD) | 2012 | ACM SIGKDD Workshop on Health Informatics (HI-KDD) |
| **f6** | Multi-source learning with block-wise missing data for Alzheimer's disease prediction | Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '13 | 2013 | ACM Press |
| **f7** | Alzheimer's Disease Risk Assessment Using Large-Scale Machine Learning Methods | PLoS ONE | 2013 | Public Library of Science |
| **f8** | Integrative analysis of multi-dimensional imaging genomics data for alzheimer's disease prediction | Frontiers in Aging Neuroscience | 2014 | Frontiers |
| **f9** | Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects | NeuroImage | 2015 | Academic Press |
| **f11** | Imaging-based enrichment criteria using deep learning algorithms for efficient clinical trials in mild cognitive impairment | Alzheimer's and Dementia | 2015 | Elsevier |
| **f13** | Alzheimer's Early Prediction with Electroencephalogram | Procedia Computer Science | 2016 | The Author(s) |
| **f14** | A Machine Intelligence Designed Bayesian Network | Procedia Computer Science | 2016 | Elsevier |

| | | | | |
|---|---|---|---|---|
| | Applied to Alzheimer's Detection Using Demographics and Speech Data | | | |
| **f16** | Cascaded Multi-view Canonical Correlation (CaMCCo) for Early Diagnosis of Alzheimer's Disease via Fusion of Clinical, Imaging and Omic Features | Scientific Reports | 2017 | Nature Publishing Group |
| **f17** | Classification of Alzheimer's disease and prediction of mild cognitive impairment-to-Alzheimer's conversion from structural magnetic resource imaging using feature ranking and a genetic algorithm | Computers in Biology and Medicine | 2017 | Pergamon |
| **f19** | Using high-dimensional machine learning methods to estimate an anatomical risk factor for Alzheimer's disease across imaging databases | NeuroImage | 2018 | Academic Press |
| **f20** | Big Data Analytical Approaches to the NACC Dataset: Aiding Preclinical Trial Enrichment | Alzheimer Disease and Associated Disorders | 2018 | NIH Public Access |
| **r21** | A Disease State Fingerprint for Evaluation of Alzheimer's Disease | Journal of Alzheimer's Disease | 2011 | NIH |
| **r22** | Alzheimer's disease biomarker discovery using in silico literature mining and clinical validation | Journal of Translational Medicine | 2012 | BioMed Central |
| **r24** | An efficient approach for differentiating Alzheimer's disease from normal elderly based on multicenter MRI using gray-level invariant features | PLoS ONE | 2014 | Public Library of Science |
| **r25** | Automated hippocampal shape analysis predicts the onset of dementia in mild cognitive impairment | NeuroImage | 2011 | Academic Press |
| **r26** | Automatically extracting sentences from Medline citations to support clinicians' information needs | Journal of the American Medical Informatics Association | 2013 | Oxford University Press |
| **r27** | Boosting power for clinical trials using classifiers based on multiple biomarkers | Neurobiology of Aging | 2010 | Elsevier |

| | | | | |
|---|---|---|---|---|
| **r28** | Building Disease-Specific Drug-Protein Connectivity Maps from Molecular Interaction Networks and PubMed Abstracts | PLoS Computational Biology | 2009 | Public Library of Science |
| **r29** | Classification and prediction of clinical diagnosis of Alzheimer's disease based on MRI and plasma measures of α-/γ-tocotrienols and γ-tocopherol | Journal of Internal Medicine | 2013 | Wiley/Blackwell (10.1111) |
| **r31** | Cognitive reserve and Alzheimer's disease biomarkers are independent determinants of cognition | Brain | 2011 | Oxford University Press |
| **r32** | Cognitive, Genetic, and Brain Perfusion Factors Associated with Four Year Incidence of Alzheimer's Disease from Mild Cognitive Impairment | Journal of Alzheimer's Disease | 2014 | Oxford University Press |
| **r33** | Compensatory mechanisms in higher-educated subjects with Alzheimer's disease: a study of 20 years of cognitive decline | Brain | 2014 | Oxford University Press |
| **r34** | Deep Learning Based Imaging Data Completion for Improved Brain Disease Diagnosis | Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014 | 2014 | Springer |
| **r35** | Depression and incident alzheimer disease: The impact of disease severity | American Journal of Geriatric Psychiatry | 2015 | Elsevier |
| **r36** | Detecting cognitive impairment by eye movement analysis using automatic classification algorithms | Journal of Neuroscience Methods | 2011 | Elsevier |
| **r37** | Detection of early Alzheimer's disease in MCI patients by the combination of MMSE and an episodic memory test | BMC Neurology | 2011 | BioMed Central |
| **r38** | Discovery and use of efficient biomarkers for objective disease state assessment in Alzheimer's disease | 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC'10 | 2010 | 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC'10 |
| **r39** | Discrimination between Alzheimer's disease and mild | Computational and mathematical methods in medicine | 2013 | Hindawi |

| | | | | |
|---|---|---|---|---|
| | cognitive impairment using SOM and PSO-SVM | | | |
| **r40** | Extension and refinement of the predictive value of different classes of markers in ADNI: Four-year follow-up data | Alzheimer's & Dementia | 2014 | Elsevier |
| **r41** | Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fMRI | NeuroImage | 2015 | Academic Press |
| **r44** | HypothesisFinder:’ A Strategy for the Detection of Speculative Statements in Scientific Text | PLoS Computational Biology | 2013 | Public Library of Science |
| **r45** | Latent information in fluency lists predicts functional decline in persons at risk for Alzheimer disease | Cortex | 2014 | Elsevier |
| **r46** | Locally linear embedding (LLE) for MRI based Alzheimer's disease classification | NeuroImage | 2013 | Academic Press |
| **r47** | Machine Learning Amplifies the Effect of Parental Family History of Alzheimer's Disease on List Learning Strategy | Journal of the International Neuropsychological Society | 2012 | Cambridge University Press |
| **r48** | Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease | NeuroImage | 2012 | Academic Press |
| **r49** | Predicting clinical scores from magnetic resonance scans in Alzheimer's disease | NeuroImage | 2010 | Academic Press |
| **r51** | Predictive markers for AD in a multi-modality framework: An analysis of MCI progression in the ADNI population | NeuroImage | 2011 | Academic Press |
| **r52** | Quantitative Evaluation of Disease Progression in a Longitudinal Mild Cognitive Impairment Cohort | Journal of Alzheimer's Disease | 2014 | NIH |
| **r53** | Single neuropsychological test scores associated with rate of cognitive decline in early Alzheimer disease | The Clinical neuropsychologist | 2014 | Routledge |
| **r54** | Systematic identification of risk factors for Alzheimer's | 18th Pacific Symposium on Biocomputing, PSB 2013 | 2013 | 18th Pacific Symposium on |

| | | | | |
|---|---|---|---|---|
| | disease through shared genetic architecture and electronic medical records | | | Biocomputing, PSB 2013 |
| **r55** | The Association of Neuropsychiatric Symptoms in MCI with Incident Dementia and Alzheimer Disease | The American Journal of Geriatric Psychiatry | 2013 | Elsevier |
| **r56** | Vascular and amyloid pathologies are independent predictors of cognitive decline in normal elderly | Brain | 2015 | Oxford University Press |
| **r57** | Antihypertensive drugs decrease risk of Alzheimer disease Ginkgo Evaluation of Memory Study | Neurology | 2013 | Wolters Kluwer Health, Inc. on behalf of the American Academy of Neurology |
| **r58** | Integrative EEG biomarkers predict progression to Alzheimer's disease at the MCI stage | Frontiers in Aging Neuroscience | 2013 | Frontiers |